Financial Services Data Management: Big Data Technology in Financial Services
Introduction: Big Data in Financial Services ......................................... 1
What is Driving Big Data Technology Adoption in Financial Services? 3
  Customer Insight .................................................................................. 3
  Regulatory Environment ........................................................................ 3
  Explosive Data Growth ......................................................................... 4
  Technology Implications ....................................................................... 4
The Big Data Technology Continuum ...................................................... 5
  Handling Unstructured Data ................................................................. 6
  Handling Semi-Structured Data ............................................................. 7
  Handling Structured Data ..................................................................... 7
  Adding New Dimensions to the Decomposition Framework ................. 8
Mapping Oracle Products to Big Data Technology Requirements ............. 9
  The Oracle Database 11g: Beyond Relational Technologies .......... 10
  Hadoop and the Oracle Big Data Appliance ........................................ 11
  Business Intelligence and Dynamic Information Discovery .............. 11
  Why Engineered Systems Matter ......................................................... 12
  Delivering Real-Time Decisions .......................................................... 14
Oracle Platform Solutions for Big Data use Cases ................................. 14
  Big Data Platform for Risk, Reporting and Analytics .......................... 14
  Platform for Data-Driven Customer Insight and Product Innovation. 16
  Platform for Security, Fraud and Investigations ................................ 18
Why Oracle ............................................................................................. 19
Introduction: Big Data in Financial Services

The Financial Services Industry is amongst the most data driven of industries. The regulatory environment that commercial banks and insurance companies operate within requires these institutions to store and analyze many years of transaction data, and the pervasiveness of electronic trading has meant that Capital Markets firms both generate and act upon hundreds of millions of market related messages every day. For the most part, financial services firms have relied on relational technologies coupled with business intelligence tools to handle this ever-increasing data and analytics burden. It is however increasingly clear that while such technologies will continue to play an integral role, new technologies – many of them developed in response to the data analytics challenges first faced in e-commerce, internet search and other industries – have a transformative role in enterprise data management.

Consider a problem faced by every top-tier global bank: In response to new regulations, banks need to have a ‘horizontal view’ of risk within their trading arms. Providing this view requires banks to integrate data from different trade capture systems, each with their own data schemas, into a central repository for positions counter-party information and trades. It’s not uncommon for traditional ETL based approaches to take several days to extract, transform, cleanse and integrate such data. Regulatory pressure however dictates that this entire process be done many times every day. Moreover, various risk scenarios need to be simulated, and it’s not uncommon for the simulations themselves to generate terabytes of additional data every day. The challenge outlined is not only one of sheer data volumes but also of data variety, and the timeliness in which such varied data needs to be aggregated and analyzed.

Now consider an opportunity that has largely remained unexploited: As data driven as financial services companies are, analysts estimate that somewhere between 80 and 90 percent of the data that banks have is unstructured, i.e., in documents and in text form. Technologies that enable businesses to marry this data with structured content present an enormous opportunity for improving business insight for financial institutions. Take for example, information stored in insurance claim systems. Much valuable information is captured in text form. The ability to parse text information and combine the extracted information with structured data in the claims...
database will not only enable a firm to provide a better customer experience, it also may enhance their fraud detection capabilities.

The above scenarios were used to illustrate a few of the challenges and potential opportunities in building a comprehensive data management vision. These and other data management related challenges and opportunities have been succinctly captured and classified by others under the ‘Four Vs’ of data – Volume, Velocity, Variety and Value.

The visionary bank needs to deliver business insights in context, on demand, and at the point of interaction by analyzing every bit of data available. Big Data technologies comprise the set of technologies that enable banks to deliver to that vision. To a large extent, these technologies are made feasible by the rising capabilities of commodity hardware, the vast improvements in storage technologies, and corresponding fall in the price of computing resources. Given that most literature on Big Data relegate established technologies such as RDBMS to the ‘has been’ heap, it is important that we stress that relational technologies continue to play a central role in data management for banks, and that Big Data technologies augment the current set of data management technologies used in banks. Later sections of this paper will expand on this thought and explain how relational technology is positioned in the Big Data technology continuum.

This paper broadly outlines Oracle’s perspective on Big Data in Financial Services starting with key industry drivers for Big Data. Big Data comprises several individual technologies, and the paper outlines a framework to uncover these component technologies, then maps those technologies to specific Oracle offerings, and concludes by outlining how Oracle solutions may address Big Data patterns in Financial Services.
What is Driving Big Data Technology Adoption in Financial Services?

There are several use cases for big data technologies in the financial services industry, and they will be referred to throughout the paper to illustrate practical applications of Big Data technologies. In this section we highlight three broad industry drivers that accelerate the need for Big Data technology in the Financial Services Industry.

Customer Insight

Up until a decade or so ago, it may be said that banks, more than any other commercial enterprise, owned the relationship with consumers. A consumer’s bank was the primary source of the consumer’s identity for all financial, and many non-financial transactions. Banks were in firm control of customer relationship, and the relationship was for all practical purposes as long-term as the bank wanted it to be. Fast forward to today, and the relationship is reversed. Consumers now have transient relationships with multiple banks: a current account at one that charges no fees, a savings accounts with a bank that offers high interest, a mortgage with a one offering the best rate, and a brokerage account at a discount brokerage. Moreover, even collectively, financial institutions no longer monopolize a consumer’s financial transactions. New entrants peer-to-peer services; and the Paypals, Amazons, Googles and Walmarts of the world – have had the effect of disintermediating the banks. Banks no longer have a complete view of their customer’s preferences, buying patterns and behaviors. This problem is exacerbated by the fact that social networks now capture very valuable psychographic information – the consumer’s interests, activities and opinions.

The implication is that even if banks manage to integrate information from their own disparate systems, which in itself amounts to a gargantuan, a fully customer-centric view may not be attained. Gaining a fuller understanding of a customer’s preferences and interests are prerequisites for ensuring that banks can address customer satisfaction and for building more extensive and complete propensity models. Banks must therefore bring in external sources of information, information that is often unstructured. Valuable customer insight may also be gleaned from customer call records, customer emails and claims data, all of which are in textual format. Bringing together transactional data in CRM systems and payments systems, and unstructured data both from within and outside the firm requires new technologies for data integration and business intelligence to augment the traditional data warehousing and analytics approach. Big Data technologies therefore play a pivotal role in enabling customer centrivity in this new reality.

Regulatory Environment

The spate of recent regulations is unprecedented for any industry. Dodd-Frank alone adds hundreds of new regulations that affect banking and securities industries. For example, these demands require liquidity planning and overall asset and liability management functions to be fundamentally rethought. Point-in-time liquidity positions currently provided by static analysis of relevant financial ratios are no longer sufficient, and a more near real-time view is being required. Efficient allocation of capital is now seen as a major competitive advantage, and risk-adjusted performance calculations require new points
of integration between risk and finance subject areas. Additionally, complex stress tests, which put enormous pressure on the underlying IT architecture, are required with increasing frequency and complexity. On the Capital Markets side, regulatory efforts are focused on getting a more accurate view of risk exposures across asset classes, lines of business and firms in order to better predict and manage systemic interplays. Many firms are also moving to a real-time monitoring of counterparty exposure, limits and other risk controls. From the front office all the way to the boardroom, everyone is keen on getting holistic views of exposures and positions and of risk-adjusted performance.

Explosive Data Growth

Perhaps the most obvious driver is that financial transaction volumes are growing leading to explosive data growth in financial services firms. In Capital Markets, the pervasiveness of electronic trading has lead to a decrease in the value of individual trades and an increase in the number of trades. The advent of high turnover, low latency trading strategies generates considerable order flow and an even larger stream of price quotes. Complex derivatives are complicated to value and require several data points to help determine, among other things, the probability of default, the value of LIBOR in the future, and the expected date of the next ‘Black Swan’ event. In addition, new market rules are forcing the OTC derivative market – the largest market by notional value – toward an electronic environment.

Data growth is not limited to capital markets businesses. The Capgemini/RBS Global Payments study for 2011 estimates that the global volume for electronic payments is about 260 billion and growing between 15 and 22% for developing countries. As devices that consumers can use to initiate core transactions proliferate, so too do the number of transactions they make. Not only is the transaction volume increasing, the data points stored for each transaction are also expanding. In order to combat fraud and to detect security breaches, weblog data from bank’s Internet channels, geospatial data from smart phone applications, etc., have to be stored and analyzed along with core operations data. Up until the recent past, fraud analysis was usually performed over a small sample of transactions, but increasingly banks are analyzing entire transaction history data sets. Similarly, the number of data points for loan portfolio evaluation is also increasing in order to accommodate better predictive modeling.

Technology Implications

The technology ramifications of the broad industry trends outlined above are:

- More data, and more different data types: Rapid growth in structured and unstructured data from both internal and external sources requires better utilization of existing technologies and new technologies to acquire, organize, integrate and analyze data.

- More change and uncertainty: Pre-defined, fixed schemas may be too restrictive when combining data from many different sources, and rapidly changing needs imply schema changes must be allowed more easily.

- More unanticipated questions: Traditional BI systems work extremely well when the questions to be asked are known. But business analysts frequently don’t know all the questions they need to ask.
Self-service ability to explore data, add new data, and construct analysis as required is an essential need for banks driven by analytics.

- More real-time analytical decisions: Whether it is a front office trader or a back office customer service rep, business users demand real-time delivery of information. Event processors, real-time decision making engines and in-memory analytical engines are crucial to meeting these demands.

**The Big Data Technology Continuum**

So how do we address the technology implications summarized in the previous section? The two dimensional matrix below provides a convenient starting, albeit incomplete, framework for decomposing the high-level technology requirements for managing Big Data. The figure below depicts, along the vertical dimension, the degree to which data is structured: Data can be unstructured, semi-structured or structured. The second dimension is the lifecycle of data: Data is first acquired and stored, then organized and finally analyzed for business insight. But before we dive into the technologies, a basic understanding of key terminology is in order.

- We define the structure in ‘structured data’ in alignment with what is expected in relational technologies – that the data may be organized into records identified by a unique key, with each record having the same number of attributes, in the same order. Because each record has the same number of attributes, the structure or schema need be defined once as metadata for the table, and the data itself need not have metadata embedded in it.

- Semi-structured data also has structure, but the structure can vary from record to record. Records in semi-structured data are sometimes referred to as jagged records because each record may have variable number of attributes and because the attributes themselves may be compound constructs, i.e. be made up of sub-attributes like in an XML document. Because of the variability in structure, metadata for semi-structured data has to be embedded within the data: for e.g., in the form of an XML schema or as name-value pairs that describe the names of attributes and their respective values, within the record. If the data contains tags or other markers to identify names and the positions of attributes within the data, the data can be parsed to extract these name-value pairs.

- By unstructured data, we mean data for which structure does not conform to the two other classifications discussed above. Strictly speaking, unstructured text data usually does have some structure -- for e.g., the text in a call center conversation record has grammatical structure -- but the structure does not follow a record layout, nor are there any embedded metadata tags describing attributes. Of course, before unstructured data can be used to yield business insights, it has to be transformed into some form of structured data. One way to extract entities and relationships from unstructured text data is by using natural language processing (NLP). NLP extracts parts of speech such as nouns, adjectives, subject-verb-object relationships; commonly identifiable things such as places, company names, countries, phone numbers, products, etc.; and can also identify and score sentiments about products, people, etc. It’s also possible to augment these processors by supplying a list of significant entities to the parser for named entity extraction.
Figure 1: Big Data Technology continuum

The diagram above might justifiably convey that a myriad of disparate technologies are needed to comprehensively handle Big Data requirements for the enterprise. However, these are not ‘either/or’ technologies. They are to be viewed as part of a data management continuum: each technology enjoys a set of distinct advantages depending on the phase in the lifecycle of data management and on the degree of structure within data it needs to handle, and so these technologies work together within the scope of an enterprise architecture.

The two points below are expanded on further along in this section, but they are called out here for emphasis:

- The diagram does not imply that all data should end up in a relational data warehouse before analysis may be performed. Data needs to be organized for analysis, but the organized data may reside on any suitable technology for analysis.

- As the diagram only uses two dimensions for decomposing the requirements, it does not provide a complete picture. For example, the diagram may imply that structured data is always best handled in a relational database. That’s not always the case, and the section on handling structured data explains what other technologies may come into play when we consider additional dimensions for analysis.

Handling Unstructured Data

Unstructured data within the bank may be in the form of claims data, customer call records, content management systems, emails and other documents. Content from external sources such as Facebook, Twitter, etc., is also unstructured. Often, it may be necessary to capture such unstructured data first before processing the data to extract meaningful content. File systems of course can handle any type of data as they simply store data. Distributed file systems are file systems architected for high
performance and scalability. They exploit parallelism that is made possible because these file systems are spread over several physical computers (from 10s to few thousand nodes). Data captured in distributed file systems must later be organized (reduced, aggregated, enriched, and converted into semi-structured or structured data) as part of the data lifecycle.

Dynamically indexing engines are relatively new class of databases in which no particular schema is enforced or defined. Instead, a ‘schema’ is dynamically built as data is ingested. In general, they work something akin to web search engines in that they crawl over data sources they are pointed at, extracting significant entities and establishing relationships between these entities using Natural Language Parsing or other text mining techniques. The extracted entities and relationships are stored as a graph structure within the database. These engines therefore simultaneously acquire and organize unstructured data.

Handling Semi-Structured Data

Semi-structured data within the bank may exist as loan contracts, in derivatives trading systems, as XML documents and HTML files, etc. Unlike unstructured data, semi-structured data contains tags to mark significant entity values contained within it. These tags and corresponding values are key-value pairs. If the data is in such a format that these key-value pairs need to be extracted from within it, it may need to be stored on a distributed file system for later parsing and extraction into key-value databases. Key-value stores are one in a family of NoSQL database technologies -- some others being graph databases and document databases – which are well suited for storing semi-structured data. Key-value stores do not generally support complex querying (joins and other such constructs) and may only support information retrieval using the primary key and in some implementations using an optional secondary key. Key-values stores like the file systems described in the previous section are also often partitioned, enabling extremely high read and write performance. But unlike in distributed file systems where data can be written and read in large blocks, key-value stores support high performance for single-record reads and writes only.

That these newer non-relational systems offer extreme scale and/or performance is accepted. But this advantage comes at a price. As data is spread across multiple nodes for parallelism there is increased likelihood of node failures, especially when cheaper commodity servers are used to reduce the overall system cost. In order to mitigate the increased risk of node failures these systems replicate data on two or often three nodes. The CAP Theorem put forward by Prof. Eric Brewer states that such systems have to choose two from among the three properties of Consistency, Availability and Partition Tolerance. And most implementations choose to sacrifice Consistency, the C in ACID, thereby redefining themselves as BASE systems (Basically Available Soft-state Eventually consistent).

Handling Structured Data

Banks have applications that generate many terabytes of structured data and have so far relied almost exclusively on relational technologies for managing this data. However, the Big Data technology movement has risen partly from the limitations of relational technology, and the most serious limitations may be summed up as:
Relational technologies were engineered to handle needs not always required. For example, relational systems can handle complex querying needs and they adhere to strict ACID properties. These capabilities are not always required, but because they are always “on”, there is an overhead associated to relational systems that sometimes constrains other more desired properties such as performance and scalability. To make the argument more concrete, let’s take an example scenario: It wouldn’t be unusual for a medium to large size bank to generate 5-6 terabytes of structured data in modeling exposure profiles of their counterparties using Monte Carlo simulations (assuming 500000 trades, 5000 Scenarios). Much more data would be generated if stress tests were also performed. What’s needed is a database technology that can handle huge data volumes with extremely fast read (by key) and write speeds. There is no need for strict ACID compliance; availability needs are less than in say, a payment transaction system; there are no complex queries to be executed against this data; and it would be more efficient for the application that generates the data (the Monte Carlo runs) to have local data storage. Although the data is structured, a relational database may not be the optimal technology here. Perhaps a NoSQL database or distributed file system or even a data grid (or some combinations of technologies) may be faster and more cost effective in this scenario.

The rigidity enforced by the relational schema model requires more upfront planning and may make the schema more brittle in the face of changing requirements. This is often felt when data from multiple sources needs to be aggregated into one database, and especially when data from source systems are in different formats.

While relational technologies may be challenged in meeting some of these demands, the model benefits tremendously from its structure. These technologies remain the best way to organize data in order to quickly and precisely answer complex business questions, especially when the universe of such questions is known. They remain the preferred technology for systems that have complex reporting needs. Also, if ACID properties and reliability are must haves for applications such as core banking and payments, few other technologies meet the demands for running their mission critical, core systems.

Moreover, many limitations of relational technology implementations like scale and performance are addressed in specific implementations of the technology, and we discuss the Oracle approach to extending the capabilities of the Oracle Database implementation – both, in terms of its ability to scale and its ability to handle different types of data -- in the next section.

Adding New Dimensions to the Decomposition Framework

In the previous sections we used the two dimensions shown in Figure 1 to uncover the technologies required to handle big data needs. In this section we outline two additional dimensions that can be used for further decomposition of technology requirements.

Handling Real-Time Needs

Real-time risk management, in-process transactional fraud prevention, security breach analytics, real-time customer care feedback and cross-selling analytics, all necessitate the acquisition, organization and analysis of large amounts of data in real-time. The rate at which data arrives and the timeliness in responding to incoming data is another dimension we may apply to the technology requirements.
decomposition framework. Acquiring data at extremely fast rates and performing analysis on such data in real-time requires different set of technologies than previously discussed.

Three most common technologies for handling real-time analytics are Complex Event Processors, in-memory Distributed Data Grids and in-memory Databases. Complex Event Processing (CEP) engines provide a container for analytical applications that work on streaming data (market data, for example). Unlike in databases, queries are evaluated against data, in-memory, and continuously as data arrives into the CEP engine. CEP engines are therefore an essential component of event-driven architectures. CEP engines have found acceptance in the front office for algorithmic trading, etc., but they have wide applicability across all lines of business and even in retail banking; for detecting events in real-time as payment or core banking transactions are generated.

Distributed data grid technologies play a complementary role to CEP engines. Data grids not only store data in memory (usually as objects, i.e., in semi-structured form) they also allow distributed processing on data in memory. More sophisticated implementations support event based triggers for processing and MapReduce style processing across nodes. Using a distributed data grid, transactions can be collected and aggregated in real-time from different systems and processing can be done on the aggregated set in real-time. For example a bank could integrate transactions from different channels in a distributed data grid and have real-time analytics run on the collected set for superior multi-channel experience.

Reducing Data for Analysis

A guiding principle at the heart of the Big Data movement is that all data is valuable and that all data must be analyzed to extract business insight. But not all data sets contain equal amount of value, which is to say that the value-density or "signal-to-noise ratio" of data sets within the bank differ. For example, data from social media feeds maybe less value-dense than data from the bank's CRM system. Value-density of data sets provides us with another dimension to apply to the framework for decomposing technology requirements for data management. For example, it may be far more cost effective to first store less value-dense data such as social media streams on a distributed file system than in a relational database. As you move from left to right in the framework, low-value density data should be aggregated, cleaned and reduced to high value-density data that is ready to be analyzed. This is not to say that the diagram implies that all data should eventually be stored in a relational data warehouse for analysis and visualization. Data can be organized in-place: The output of a MapReduce process on Hadoop may be the HDFS file system itself and analytical tools for Big Data should be able to provide present outputs from analysis of data present in both non-relational and relational systems.

Mapping Oracle Products to Big Data Technology Requirements

Figure 2 below maps Oracle’s engineered systems and software products to components in the technology requirements decomposition matrix shown the previous section. As the figure shows, Oracle offers a comprehensive portfolio of products to cover the Big Data technology continuum. As more detailed and thorough product descriptions are available in whitepapers and product sheets on Oracle websites, no attempt is made to duplicate that information in this section. Instead this section highlights key product related topics that showcase Oracle's approach to Big Data.
Although all Oracle software will run on standalone hardware servers from Oracle or from any of the leading hardware vendors, Oracle’s engineered systems approach offers a set of unique and differentiated values. Engineered systems are combinations of hardware and software, with the software optimized and tuned to exploit the specific hardware architecture found in each engineered system. It should be noted that the external interfaces to all Oracle software remains unchanged and it’s only the components that are never exposed to consuming applications that are modified in the optimization process. This ensures that there is no vendor lock-in simply because engineered systems are used. Referring to the diagram, the engineered systems shown are Oracle Big Data Appliance, Oracle Exalytics, Oracle Exalogic and Oracle Exadata. The software components shown in grey rounded rectangles are grouped by the engineered system they are optimized to run on. In a nutshell, Oracle Exadata is an engineered system for Oracle Database; Oracle Exalogic provides highly optimized Java middleware for applications; Oracle Exalytics is for ultra fast analytics and visualization; and Oracle Big Data Appliance is for acquiring and processing massive data sets using Cloudera’s distribution of Apache Hadoop, Oracle NoSQL Database and the open source R statistical analysis platform.

Figure 2: Oracle products for Big Data Management

The Oracle Database 11g: Beyond Relational Technologies

Oracle Database may be best known for its relational database technology, but the database has many additional capabilities that make it particularly well suited in the Big Data space. The Advanced Analytics option of the database ships with the Oracle R Enterprise, which allows analysts to build analytical applications using the popular open source R programming environment. Oracle R Enterprise is embedded within the database engine, and allows statistical and predictive analytical models written in R to access data in Oracle tables without having to write SQL queries. These analytical applications may also access external data held in Hadoop via the Oracle Direct Connector for HDFS. Oracle Database’s embedded analytical capability is not limited to R: it’s augmented with
Text mining and Data mining modules. These options allow users to create predictive models, text analysis applications that run within the database. Oracle Data Mining, a component of the Oracle Advanced Analytics option, is a collection of machine learning algorithms embedded in Oracle Database, which allows the discovery of new insights, segments and associations from existing data.

When used with Oracle Exadata, parallel execution, enhanced scalability and performance are automatically available to these analytical applications. But whether using engineered systems or not, the concept of performing analytics right in the database is a very powerful one, and follows the Big Data philosophy of moving computation to the data rather than the other way around. The combination of high throughput analytics with engineered systems has enabled several financial firms to dramatically reduce the time it takes to run analytical workloads. Whether it's EOD batch processing, on-demand risk calculation, or pricing and valuations, firms are able to do a lot more in much less time, directly affecting the business by enabling continuous, on-demand data processing.

Hadoop and the Oracle Big Data Appliance

Hadoop is a complete, open-source stack for storing and analyzing massive amounts of data, with several distributions available. Oracle has partnered with Cloudera, the leader in Apache Hadoop-based distributions.

The Oracle Big Data Appliance is an engineered system optimized to run Big Data workloads using Cloudera's distribution including Apache Hadoop, Oracle NoSQL DB and R. It offers unsurpassed performance, as well as pre-integrated and pre-configured application and hardware platform components, which are all engineered to work together (which no small feat in the case of Hadoop). It uses an Infiniband backbone to connect the cluster nodes together, enabling superior MapReduce performance without the need to setup and administer an Infiniband network.

Business Intelligence and Dynamic Information Discovery

The architectural approach of having the Enterprise Data Warehouse (EDW) as the single source of truth for business decision making is widely subscribed to because of its conceptual elegance and simplicity. Traditional business intelligence tools (like Oracle Business Intelligence Enterprise Edition) are used in conjunction with the ‘EDW/Datamart’ approach. These BI tools are excellent at answering questions that are known and have been modeled for in the Datamarts.

However, EDW and BI tools approach was predicated on the belief that all analytical questions could be modeled at design time. Datamarts are designed to answer specific and known questions. New questions, previously not modeled for, cannot easily be answered. Moreover, while the boundary between operational tasks and decision making tasks may have been clear in the past, today this boundary is increasingly blurred: Real-time decision making (for e.g., presenting contextual offers to a customer during a transactional interaction) requires simultaneous access of data in both operational and analytical data stores. Given these limitations, traditional ways of gaining business insight need to be complemented with a form of dynamic information discovery over data from all sources within the bank. Dynamic information discovery is necessary not only to accommodate BI reporting across different sources and types of data, but also because not all questions are known at design time. The business user can’t be sure which questions will matter, perhaps because the business environment is
prone to unanticipated changes, or simply because there are too many variables to work out all the combinations of questions.

Oracle Endeca Information Discovery (EID) allows dynamic discovery of information across structured and unstructured sources of information. EID is enabled by the Oracle Endeca server, which automatically organizes information for search, discovery & analysis from documents, content management systems, data warehouses and operational data stores without the need for predefining a model. EID is a complementary technology for business intelligence to Oracle’s Business Intelligence Foundation Suite tools, and provides a search based guided navigation UI for information discovery. The capabilities of these software applications are further enhanced when run within the Oracle Exalytics engineered machine platform. Oracle Exalytics integrates a purpose built column-oriented version of Oracle’s TimesTen in-memory database, Oracle Business Intelligence Foundation Suite software and high memory capacity hardware, and the BI software intelligently caches information based on knowledge about data access patterns within the in-memory database.

Why Engineered Systems Matter

Parallelism and data locality aware processing are two techniques exploited by newer technologies to provide massive scale and high performance. Parallelism in Hadoop is employed for file access in HDFS and also in the MapReduce framework for processing. Data locality for processing means that processing is done as close to the data as possible, i.e., instead of data being moved to nodes that do processing, the processing is moved to nodes that have the data, thus ensuring that vast data sets are not moved over the network. In Hadoop, the Map process is done with data locality. A similar technique of offloading the database processing to the storage layer where the data is located is employed in Oracle Exadata, giving Exadata several fold scale and performance advantage in comparison to other database solutions. This is possible because Oracle has engineered the entire stack and optimized it for extreme performance. For example, within an Exadata environment, the Oracle database software recognizes that the storage within the environment isn’t general purpose SAN storage. With this knowledge and using a unique feature of Oracle Exadata called Exadata Smart Scan the database off loads queries to the storage subsystem, which performs data-local processing. Parallelism is also employed both within the database software and within storage subsystem: a full rack Exadata comes with fourteen storage nodes, which is expandable further. The net effect is that these systems break the scale and performance barriers that once limited relational technology implementations.

Take another example. In the previous section we explained that many non-relational technologies achieve scale and performance by sacrificing consistency (BASE systems). However, when the Oracle NoSQL database runs on the Oracle Big Data Appliance, this limitation can be reduced. The Big Data Appliance is an engineered system that provides 18 processing nodes (216 cores total) with 24 terabytes of local storage per node, connected via a high-speed Infiniband fabric. The high-speed fabric between nodes drastically lowers the latency of NoSQL partition replication operations. With this much reduced latency penalty, administrators may set up the Oracle NoSQL Database to synchronously replicate data to partitions providing better consistency models. Hadoop jobs also benefit from this network fabric: Although the Map phase is performed local to the data, the output
data from the map phase has to be moved over the network for the reduce phase, and the Infiniband network within the Big Data Appliance reduces the bottleneck in network latency and bandwidth in this phase of processing.

Performance and scale are excellent reasons in themselves for considering engineered systems, but there is an often overlooked, and arguably more important reason relevant to large enterprises such as banks. It is no simple task to set up a Hadoop cluster. In addition, the maintenance of large clusters of machines -- ensuring that software, OS, drivers, etc., are patched and upgraded in unison -- poses a significant cost and skill challenge to the bank’s IT department. Engineered systems allow IT to get up and running quickly and then focus more on delivering services to the business – be they infrastructure services such as MapReduce or database services or application services like sentiment analysis – and less on managing a complex infrastructure. Oracle engineered systems are developed specifically in support of this goal.

**Engineered Systems Working in Concert**

Oracle engineered systems, by definition, are engineered to run specific software components, but they are also engineered to work together in concert.

![Oracle’s Engineered Systems work in concert](image)

Software connectors optimized to work over the high-speed network interconnects enable engineered systems to work together. For example, the Big Data Appliance can be connected over Infiniband to Oracle Exadata, which in-turn may be connected to Oracle Exalytics In-Memory Machine. As each machine is optimized to handle data for a particular phase in the data management lifecycle, together they provide for the complete needs for enterprise Big Data management.

Oracle provides specific connectors and optimizations to exploit inter-system interconnectivity. In order to efficiently load data processed in HDFS on Oracle Big Data Appliance into the Oracle Database, the Oracle Loader for Hadoop (OLH) may be used as the last step in a MapReduce process. OLH itself uses MapReduce processing and creates Oracle internal format, optimized data sets. These OLH output data sets may be efficiently loaded into an Oracle Database using the Oracle Data Integrator Application Adapter for Hadoop, which simplifies data integration from Hadoop using the
Oracle Data Integrator's easy to use interface. Alternatively, instead of moving these data sets from the Big Data Appliance into Oracle Exadata, they may be accessed by SQL analytical queries running within the database on Exadata using the Oracle Direct Connector for HDFS. This connector works by creating external tables pointing at the data sets created by OLH on the Big Data appliance, allowing database tables to be joined with these external data files in SQL queries.

Delivering Real-Time Decisions

The Oracle Real-Time Decision product (RTD) is a general-purpose enterprise decision engine that is well suited to address requirements of Big Data. The platform offers a business framework that combines predictive analytics and business rules to power analytical solutions for enterprise decision management.

- From an analysis perspective, RTD can automatically build thousands of predictive models that correlate broad vectors of information with a set of declared business events. The lifecycle of those predictive models from creation to execution is fully managed by RTD and in particular those models are incrementally adjusted over time with new data coming in.

- From an execution perspective, those predictive models can in turn be used as part of the highly scalable RTD Decision Engine to express and execute channel agnostic logic combining business rules and predictive models to drive transactional process optimization. As part of this logic, the RTD self-learning, closed-loop predictive models can be combined with other Big Data scoring algorithms such as Mahoot, R or Oracle Data Mining whether in batch or real-time to operate the multi-dimensional scoring logic of RTD.

Oracle Platform Solutions for Big Data use Cases

The technologies and products described in previous sections are brought together and grouped into Oracle’s big data platform solutions. More detailed reference architectures and platform best practices will be published separately for each solution. This section provides an overview of each platform and maps a set of use cases to each solution area.

Given the broader view of Big Data that we have adopted in the paper – that Big Data is not only about handling unstructured data – not all use cases are about enabling banks to do new things. As such, many use cases listed below illustrate how big data technologies enable banks to do what was previously done better, i.e., faster, cheaper, or more efficiently.

Big Data Platform for Risk, Reporting and Analytics

‘On-Demand’ risk, especially at the desk level is now the desired goal of global banks. The objective is not only faster measurement and reporting of risk, but also measurement across asset classes. Aggregation of global positions, pricing calculations, and VaR fall within the realm of Big Data not only due to the mounting pressure to speed these calculations up well beyond the capacity of current systems, but also because of the need to deal with ever growing volumes of data. While firms have adopted the use of compute grid technologies to enable faster risk computations, the feeding of data into compute grids has been a challenge. Technologies such as Oracle Coherence complement
compute grid technologies, and enable faster calculations by allowing real time access to in-memory positions data and by allowing MapReduce style parallelization. Such an architecture helped a global bank reduce VaR calculation time from 15 hours on a 200 node grid to 20 minutes on a 60 node grid, while at the same time the number of simulations was increased almost 25 fold. An emerging use of Hadoop is in the storing the results of these computational runs. Engineered systems like the Oracle Big Data Appliance with massive amounts of local storage and optimized database technologies such as Oracle NoSQL Database and Hadoop Hive may also enable fast local storage and retrieval of data from simulation runs, which may run locally on the appliance.

Solution Components:
- Desk level, real-time risk aggregation using Oracle Coherence Grid on Oracle server grid.
- Oracle Big Data Appliance used to parallelize ETL for data in flat files and/or semi-structured trade data.
- Risk Warehouse on Oracle Exadata with in-database analytics. Oracle Data Integrator and Oracle GoldenGate for ETL from transactional systems.
- Oracle Business Intelligence Foundation Suite running on Oracle Exalytics produces dashboards and reports, aggregated at various levels.

For Enterprise Risk Management, the added challenge is of data integration from many more disparate systems. It’s not uncommon for data to be sent from source systems as flat files. The Oracle Big Data Appliance and Oracle Big Data Connectors enable ETL style workflow process to be parallelized on Hadoop before the processed data is loaded into an enterprise risk warehouse. For the risk warehouse itself, Oracle Exadata with in-database analytics virtually eliminates the performance bottlenecks typically associated with running SQL processes on database servers. Enterprise risk software,
including Oracle Financial Services Liquidity Risk Management, benefit from this capability. In a benchmark Oracle Financial Services Liquidity Risk Management running on Oracle Exadata Database Machine calculated business-as-usual liquidity gaps for 371 million cash flows across 66 million accounts in just 69 minutes. After applying modified behavior assumptions to simulate adverse market conditions, stressed liquidity gaps were calculated in only 10 minutes. With the ability to execute an individual stress test run in mere minutes, institutions can refine their scenarios to simulate any impact on business-as-usual liquidity gaps and immediately assess the effects of a given counterbalancing strategy.

**Post-Trade Analytics**

Execution algorithms are the profit driver for execution-based businesses and they are also offered as products by the sell-side to asset managers and institutional funds. Transaction Cost Analysis, which measures actual order execution performance against established benchmarks metrics, is an excellent use case for Big Data technology. TCA, initially adopted mainly as a ‘check the box’ tool for compliance with best execution regulations, has now found wide acceptance outside the compliance use. TCA is now used to assess broker performance internally, to identify outlier trades and to measure performance of algorithms sold by the sell-side. Trades from different desks can be captured on the Oracle Big Data Appliance for analysis using MapReduce and R. Processed results can then be loaded into a trade data warehouse housed on the Oracle Exadata platform. A number of large global banks have implemented trade data warehouses on Oracle Exadata and the Oracle Big Data Appliance (BDA) is a new technology that can complement Exadata. Using the BDA will allow faster trade capture on HDFS, and faster processing (using Hadoop and R) before processed data is loaded into Exadata for analysis.

Platform for Data-Driven Customer Insight and Product Innovation.
Solution Components:

- Real-time ‘decisioning’ delivered by Oracle applications running on Oracle Exalogic using underlying infrastructure consisting of Oracle CEP, Coherence, and Oracle Real-Time Decisions.
- Dynamic discovery enabled via content acquisition from all internal sources: Using Oracle Endeca Information Discovery and Oracle Endeca server with Oracle Exalytics engineered machine.
- Oracle Big Data Appliance and Hadoop processes external unstructured information before loading into customer warehouse running on Exadata; and Oracle Business Intelligence Foundation Suite running on Oracle Exalytics produces dashboards and reports from the warehouse.

Customer Profiling, Targeting and Cross-selling

Banks are looking at ways to offer new, data-driven targeted services to their customers, in order to increase revenue, reduce churn and increase customer engagement. This often requires integration of external sources of information to detect life stage events (birthdays, anniversaries, graduation dates, etc). The availability of troves of web data about almost any individual - including spending habits, risky behavior, etc. - provides valuable information that can target service offerings with a great level of sophistication. Add location information (available from almost every cell phone) and you can achieve almost surgical customer targeting.

Integrating external customer data requires identity resolution engines that use multiple factors to match the identity of an external individual (facebook account, etc) with that of the bank’s customer. Targeted real-time offers and discounts requires integration of transactions across various channels in real-time, coupled with real-time decisioning software.

Oracle has a number of market leading technologies that enable banks to implement an optimized cross-channel strategy. Oracle worked with a large US Bank that handles over 400 million interactions across channels each month to implement a customized customer treatment model. The solution utilized Oracle Coherence for the underlying transaction integration and Oracle Real-Time Decisions software to generate recommendations, which are pushed in real-time to smartphones, Facebook, website and call-center channels.

Product Innovation

Retail Banks, Investment Banks and Insurance companies are looking for new methods and techniques to develop new products and trading strategies. Retail customers are more likely to be attracted and retained with personalized banking products - hence, lifetime value goes up. Developing and launching new innovative products first in the market is a key business strategy for credit card and debit card product lines. Additionally, Insurance insurances companies need to grow their market share by tailoring insurance policies for each customer, using fine-grained analytics.

The challenge is current technology is just not suited to support the ultimate needs for product innovation that must consider data found outside traditional transaction systems and systems of record such as customer interaction weblogs, social media and news. The challenge is how to store and analyze this massive amount of unstructured data in an efficient and cost effective way to better managed portfolios.
Sentiment Analysis and Brand Reputation

Whether looking for broad economic indicators, market indicators, or sentiments concerning a specific organization or its stocks, there is obviously a trove of data on the web to be harvested, available from traditional as well as social media sources. While keyword analysis and entity extraction have been with us for a while, and are available from several data vendors, the availability of social media sources is relatively new, and has certainly captured the attention of many people looking to gauge public sentiment.

Sentiment analysis can be considered straightforward, as the data resides outside the firm and is therefore not bound by organizational boundaries. In fact, sentiment analysis is becoming so popular that a couple of hedge funds are basing their entire strategies on trading signals generated by tweeter feed analytics. While this is an extreme example, most firms at this point are using sentiment analysis to gauge public opinion about specific companies, markets or the economy as a whole.

Financial Services companies want to ensure their company’s reputation is protected by screening potential and existing customers for social, environment and ethical lapses, and a few banks have started to use Big Data to analyze a customer’s commercial activities to determine risk potential. Analyzing social media, news, financial reports maybe a preventive measure against doing business with people and organizations that could produce bad publicity.

Platform for Security, Fraud and Investigations

Fraud Detection

Banks and Insurance companies are looking to gain deeper insight into their customer interactions across all channels to reduce potential losses from collection recovery situations and claims processing. Customer treatments programs can be aligned to each customer’s situation to execute the most effective way to recover unpaid fees and outstanding loans and thereby reduce losses.

Fraud detection is also related, in as much as a similar point can be made: i.e., that correlating data from multiple, unrelated sources has the potential to catch more fraudulent activities earlier than current methods. Consider for instance the potential of correlating Point of Sale data (available to any credit card issuer) with web behavior analysis (either on the bank’s site or externally), and potentially with other financial institutions or service providers such as First Data or SWIFT, to detect suspect activities.

Payment providers have developed fraud detection tools that depend on massive datasets containing not only financial details for transactions, but IP addresses, browser information, and other technical data that will help these companies refine models to predict, identify, and prevent fraudulent activity. FIs are also considering using Big Data solutions to leverage publicly available data such as Property Deed and Assessments, to identify mortgage fraud such as equity stripping schemes.

This would go above and beyond the current Know Your Customer initiatives, watchlist screening, and the application of fundamental rules. Correlating heterogeneous data sets has the potential to dramatically improve fraud detection, and could also significantly decrease the number of false positives.
Solution components:

- Hadoop HDFS for storing related records (email, text for instigation); Hadoop MapReduce for filtering required data and NLP; Oracle NoSQL for linking unstructured and transactional data (Card Management; Payments, POS and device data for Banking & OMS/EMS/Books and records for Capital Markets).

- Oracle Data Integrator or Oracle GoldenGate and Oracle’s Big Data Connectors for relevant aggregation into Data warehouse on Oracle Exadata; Oracle Advanced Analytics option for statistical and lexical analysis.

- Oracle Business Intelligence Foundation Suite running on Oracle Exalytics produces dashboards and reports.

Reputation, Rogue Trading Detection and Investigation (eDiscovery)

Deep analytics that correlate accounting data with position tracking and order management systems can provide valuable insights that are not available using traditional data management tools. In couple of well-known cases, inconsistencies between data managed by different systems could have raised red flags if found early on, and might have prevented at least part of the huge losses incurred by the affected firms. There is also regulatory need to retrieve and respond quickly to document requests (emails, trade logs, contract documentation). From a technology perspective, the need is to have systems that enable analysis and correlation of data from application logs, emails and other activity logs in order to pull together data needed during an investigation.

Why Oracle

Acquisition, organization and analysis of Big Data today requires a whole suite of disparate products – traditional products such as Relational Databases, ETL tools, CEP engines, BI tools; and newer
products such as NoSQL Databases, Hadoop and Natural Language Processing tools. Oracle is an
accepted market leader in the traditional database/ BI product market with nearly half of the relational
database market and the largest share of the BI market. A lesser known fact is that Oracle also
develops the most widely adopted NoSQL key-value database – Berkeley DB. And with key
acquisitions of technologies such as Endeca, Oracle now has leading technologies that extend data
management and business intelligence capability beyond the traditional RDBMS / BI space.

With the acquisition of Sun Microsystems, Oracle has increasingly focused on developing engineered
systems, which are tightly integrated stacks of hardware and software optimized to run specific
workloads. New integrated machines such as the Oracle Big Data Appliance and Oracle Exalytics In-
Memory Machine are specifically developed as a part of the Big Data Technology continuum.
Importantly, these machines are engineered to work in concert with Oracle Exadata; Oracle has
created near seamless integration between these seemingly different worlds of structured and
unstructured data with tools such as Oracle Data Integrator and Oracle Loader for Hadoop.

But above all, Oracle is keenly focused on Financial Services as an industry and has market-leading
applications for Core banking, Enterprise Risk & Compliance, Enterprise Fraud, and Financial
Analytics. Oracle’s use of Big Data tools to help solve these critical issues facing the industry is where
Oracle distinguishes itself.

Oracle is doing the hard work in bringing the right technologies to solve specific business issues, and
significantly, the Oracle approach allows banks to leverage existing investment in tools and
technologies and does not require a rip-and-replace strategy. The new world of Big Data should
seamlessly integrate with the traditional world of Data Warehousing and BI, Oracle is at the forefront,
leading the charge.