ORACLE BIG DATA

Oracle Big Data Service

Bring your Present, Build your Future

Oracle Big Data Service is an automated cloud platform service designed for a diverse set of big data use cases and workloads. From agile, short-lived clusters used to tackle specific tasks to long-lived clusters that manage large data lakes, Big Data Service scales to meet all big data requirements at a low cost and with the highest levels of security. Create new big data clusters or efficiently extend your on-premise big data solutions – and leverage the full capabilities of Cloudera Enterprise Data Hub along with Oracle Big Data analytics capabilities. Take advantage of Oracle Cloud SQL to enable new and existing applications to gain insights from data across the big data landscape using Oracle's advanced SQL dialect – including data sourced from Hadoop, Object Storage, NoSQL and Kafka. And, use the languages of your choice – including Python, Scala, R and more – for machine learning, graph and spatial analytics.

Oracle Big Data Service provides a comprehensive environment for accelerating big data analytics

- Choice of Cloudera Enterprise versions
- Query and correlate big data sources with Oracle Cloud SQL
- Analyze data at scale using Oracle Machine Learning for Spark
- Use cutting edge analytics on places and networks using Oracle Big Data Spatial and Graph
- Orchestrate ELT using Oracle Data Integrator using visual code-building
- Oracle Cloud Infrastructure native big data service

BIG DATA SERVICE – OVERVIEW

Oracle Big Data Service is an automated cloud service for big data processing. It is an optimal platform to run a diverse set of workloads – from Hadoop-only processes (ETL, Spark, Hive, Impala, etc.) to interactive, all-encompassing SQL queries using Oracle Cloud SQL.

Oracle Big Data Service enables you to deliver innovation faster than ever before and with the scale and reliability you expect from Oracle, including:

- A range of Oracle Cloud Infrastructure compute options – encompassing powerful bare metal instances and flexible VM shapes
- Multiple storage options to match your functional and budgetary requirements. Achieve outstanding IO performance through the use of direct attached NVMe-based storage. Use Block Storage as a cost-effective means to reliably store data while maintain good performance. Archive data to Oracle Object Storage for the most cost-efficient, reliable data durability
- Cloudera's comprehensive software suite including Cloudera Distribution including Apache Hadoop and Apache Spark, Apache Kafka, and Cloudera Manager
- Oracle Big Data Spatial and Graph – a suite that provides cutting-edge tools for exploring and analyzing massive graphs and geo-locational data

ORACLE

- Oracle Machine Learning for Spark – which offers data scientists a familiar R interface to use distributed machine learning algorithms for data preparation, data exploration, and statistical analysis at scale
- Fast provisioning, automatically creating highly available and fully secured clusters in tens of minutes
- Simplified lifecycle operations for cluster provisioning and expansion using both a service console and API.

Big Data Service includes the best of Cloudera and Oracle big data capabilities. The service delivers an optimized deployment of Cloudera Enterprise (Data Hub Edition) – allowing customers to take advantage of all the open source features including HDFS, Spark, Hive, Impala, Solr, Kafka, and more. To enhance Spark processing, Oracle has tuned the configuration and resource management based on significant research using real world customer workloads and configurations. Oracle's optimized configuration has yielded performance gains up to 70x over existing customer configurations.

With Oracle Cloud Platform, you get the best possible cloud environment for Big Data workloads. Big Data Service complements OCI Data Flow – a fully managed Oracle Platform Service for Spark processing against the data lake. Use Big Data Service when your solution requires the complete Hadoop stack, or take advantage of OCI Data Flow for Spark-only tasks.

SECURE AND HIGHLY AVAILABLE DEPLOYMENT WITH A SINGLE CLICK

Big Data Service eases and accelerates deployments for every Hadoop cluster. By specifying a few simple options, a highly secure and optimally configured cluster is quickly provisioned.

Big Data Service automatically creates robust and highly available clusters, taking advantage of Oracle Cloud Infrastructure high availability building blocks, including redundancy, monitoring and failover (more information about Oracle Cloud Infrastructure High Availability found here). Hadoop clusters are protected against data center failures by deploying nodes across multiple Oracle Cloud Infrastructure availability domains and fault domains. In addition, clusters are protected against software service and infrastructure (e.g. Kerberos KDC and MySQL Database) failures by implementing automatic failover where possible and replication in other instances.

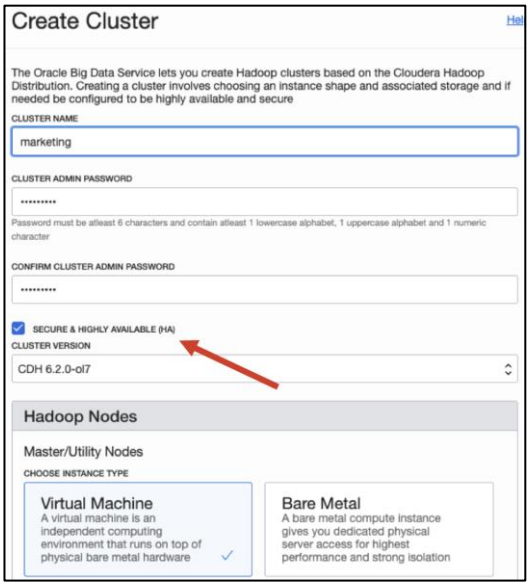

Figure 1. Create a highly secure and available cluster with a single-click.

Key Business Benefits

Oracle Big Data Service speeds time to value by providing a cloud-based automated service for big data processing using Cloudera CDH and Apache Spark. Big Data Service provides flexible compute shapes, rich security and high availability.

- Highly secure and highly available deployment with a single click
- Right size your cluster – from small development/POC clusters to comprehensive data lakes
- Workload portability between on-premises and cloud helps maximize your big data investment
- Bring the insights of big data to your existing applications via Oracle SQL

Highest Levels of Security

Securing data is critical to all enterprises. Similar to high availability features, Big Data Service leverages the rich security capabilities of Oracle Cloud Infrastructure. This includes customer isolation, segregation of operational responsibilities, visibility and audit, network security, physical security and more (see Oracle Cloud Infrastructure Security for details). But, adherence to the underlying security infrastructure is not enough. Big Data Service delivers the most secure Hadoop cluster deployment possible. And, it makes the specification of highly secure clusters simple.

Big Data Service provides strong authentication, authorization, and auditing of data in Hadoop with just a single click. It leverages all of the Cloudera distribution's data security features. Strong authentication is provided using Kerberos, which authenticates every user and ensures that rogue services are not impersonating legitimate ones.

Big Data Service leverages Apache Sentry to authorize SQL access via tools like Oracle Cloud SQL, Hive, and Impala. Sentry provides role-based access down to the table column level – ensuring that only those with the appropriate access privileges can view and analyze sensitive data.

Both data-at-rest and network encryption are capabilities included with Oracle Big Data Service. Network encryption prevents network sniffing from capturing protected data.

Finally, every operation on the Hadoop cluster is audited and tracked using Cloudera Navigator. This includes attempted file access, job executions, data definition operations, granting of roles and permissions, etc. An intuitive user interface is provided to report on the audit activity. APIs are available to integrate this audit data with third party auditing frameworks.

RIGHT SIZE CLUSTER TO MATCH YOUR WORKLOADS

Big Data Service supports a wide range of compute shapes and storage options to support any workload requirement. You can build your cluster tailored for your needs, to provide the optimal balance of costs and performance. Listed below are some examples:

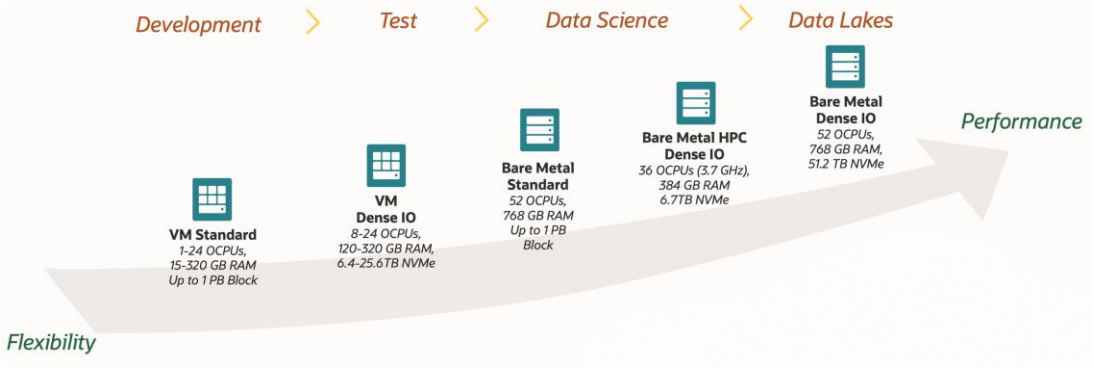

Figure 2 – Use a variety of shapes to match your requirements

- Use small VM shapes to support development and proof of concepts
- Target Bare Metal shapes for active, comprehensive data lakes, direct attached NvME storage combined with high number of cores, and memory supporting high volume workloads requiring Apache Spark based in-memory analytics, relational queries, data processing tasks, and more
- Spin up a data lab using Bare Metal servers with block storage – combining powerful servers required for analytics with flexible block storage

Oracle Big Data Service

Big Data Service is an integrated component of Oracle's comprehensive data and analytics cloud platform

Related Products

- Oracle Cloud SQL
- Oracle Autonomous Database
- Oracle Analytics Cloud
- Oracle Cloud Infrastructure Data Science
- Oracle Cloud Infrastructure Data Flow
- Oracle Cloud Infrastructure Data Catalog
- Oracle Cloud Infrastructure Storage
- Oracle Data Integrator
- Oracle Big Data Spatial and Graph
- Oracle Machine Learning for Spark
- Oracle Big Data Connectors

Start small and grow your solution. Big Data Service simplifies cluster expansion – from both a compute and storage perspective.

Flexibility in shapes extends to the structure of the cluster– enabling optimized deployment of Hadoop services to match the compute resources. A Big Data Service cluster will utilize different shapes for the roles played by the hosts. Critical nodes running master services do not have the same requirements as worker nodes that are executing tasks at scale. You simply select the compute shape for each type of node to match your workload requirement.

WORKLOAD PORTABILITY: MAXIMIZE YOUR BIG DATA INVESTMENT

Cloud may be the future of enterprise computing, but the reality is that most organizations will need to maintain a mix of public cloud, local cloud, and traditional on-premises computing for the foreseeable future. When you are building your solutions, you need to be able to deploy where it makes sense, anywhere in those three different environments. Big Data Service supports this need extremely well.

Oracle Big Data Appliance is a Cloudera-based engineered system deployed at hundreds of customers. And, of course, there are many Cloudera DIY clusters deployed across organizations. Big Data Service has been developed to work as a “cloud also” environment for these deployments. It allows you to run the same Hadoop distribution and version in the cloud that you are running in other environments. This means the same APIs, the same query engines, the same Apache Spark version, the same Hadoop management platform, etc. are used regardless of where the cluster is running. This enables workload portability – allowing you to cost effectively run your environment in the best possible location. And, there is no need to retrain administrators and developers. Examples of how you may want to leverage this capability include:

- Spin up a development cluster using a newer Cloudera version on Big Data Service and test out the latest features and potential application updates to your on-premise cluster
- Move your test environment to Big Data Service for your monthly sprint – matching your on-premise Cloudera version with cloud. Simply terminate the cluster when testing is complete.
- Migrate your data lake to Big Data Service – preserving your data and application code. Turn your on-premise cluster into a development environment.

Of course, the creation of these environments can be completely automated using Oracle Cloud Infrastructure APIs. This workload portability allows you to maximize your big data investment – taking advantage of the benefits of both on-premise and cloud deployments.

QUERY ALL YOUR DATA USING CLOUD SQL

Oracle Cloud SQL is an additionally licensed service that allows you to run Oracle SQL queries that correlate information from multiple big data sources. Because the Cloud SQL Query Server is an Oracle query engine – it supports the same queries that you would run against any other Oracle Database. This means that your existing Oracle-based applications can query Hadoop, Kafka, object storage (Oracle Object Storage and Amazon S3) without making any changes; the queries simply work. And, the queries are fast as they leverage the compute power of the cluster nodes to process data.

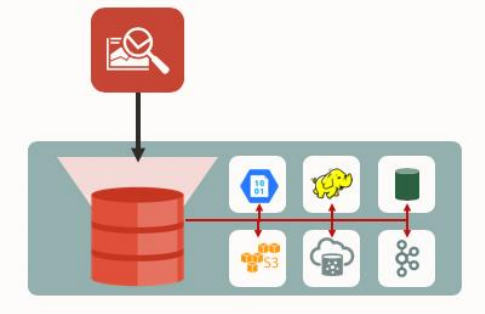

Figure 3. Use Cloud SQL to Query Hadoop, Object Storage, Kafka and NoSQL

You can easily add Cloud SQL to your Big Data Service cluster. Query Server is deployed to its own dedicated node on the cluster using a variety of shapes to match your workload – while the Cloud SQL cells are deployed to each worker node of the cluster. This deployment ensures maximum performance thru scale-out, distributed, data local processing (applying functions, filtering and aggregating data) – while isolating the Query Server execution engine to its own server.

Cloud SQL uses Hive metadata – simplifying data administration. Hive databases and tables appear as schemas and tables in the Query Server. Additional tables may be created over HDFS, object storage and Kafka sources. Data authorization leverages Apace Sentry security policies – the same policies used to authorize access to Hive metadata. There is no need to replicate security policies in the Query Server.

USE ORACLE ANALYTIC AND DATA CONNECTIVITY CAPABILITIES

Big Data Service includes licenses for Oracle's big data connectivity and analysis capabilities.

Oracle Machine Learning for Apache Spark

Oracle Machine Learning for Spark (OML4Spark) provides data scientists and business analysts an R interface for manipulating and analyzing data stored in Hadoop (Hive, Impala, Spark DataFrames, and HDFS), Oracle Database, and other JDBC sources. OML4Spark takes advantage of all the worker nodes in a Big Data Service cluster for scalable, high performance machine learning modeling. OML4Spark machine learning algorithms use the expressive R formula object optimized for Spark parallel execution.

OML4Spark brings custom Linear Model (LM), Generalized Linear Model (GLM), and MLP Neural Networks algorithms using Apache Spark execution. These algorithms have been tuned to outperform and offer better scalability than similar capabilities in Apache Spark ML. OML4Spark also provides interfaces to Apache SparkML algorithms – allowing you to choose from a variety of algorithms.

Oracle Big Data Spatial and Graph

Oracle Big Data Spatial and Graph provides advanced spatial analytic capabilities and a graph database.

The property graph component gives users a scalable graph database with industry-leading in-memory analytics. It includes 35 pre-built graph analytics enabling users to easily discover relationships, communities, influencers, and other graph patterns. The graph database is hosted on either Apache HBase or Oracle NoSQL Database and supports popular scripting languages like Python, Groovy, the open source Tinkerpop stack as well as a Java API.

The spatial analytics and services include a data enrichment service to harmonize data based on locations and place names and a wide range of 2D, 3D and raster algorithms to analyze location

relationships among persons and assets in for example social media or log data. It can apply city, state, and country categorization and process and visualize geospatial map data and satellite imagery.

Oracle Big Data Connectors

Oracle Big Data Connectors simplify data integration and analytics. Use Oracle Data Integrator to define and orchestrate complex ELT workflows. The connectors also provide high-speed loading of data in Hadoop to Oracle Autonomous Database, Oracle Exadata Cloud Service and other Oracle Database Cloud Services – with data transfer rates exceeding multiple terabytes per hour.

INTEGRATED WITH ORACLE CLOUD PLATFORM SERVICES

Big Data Service is part of a comprehensive PaaS data and analytics platform that provides a vast array of services to meet the data platform and analytics needs of an enterprise. This includes Oracle Autonomous Database, Oracle Analytics Cloud, Oracle Cloud Infrastructure Data Flow, Oracle Cloud Infrastructure Data Science and Oracle Cloud Infrastructure Data Catalog. All of these services are designed to work together, with common infrastructure and simplified connectivity.

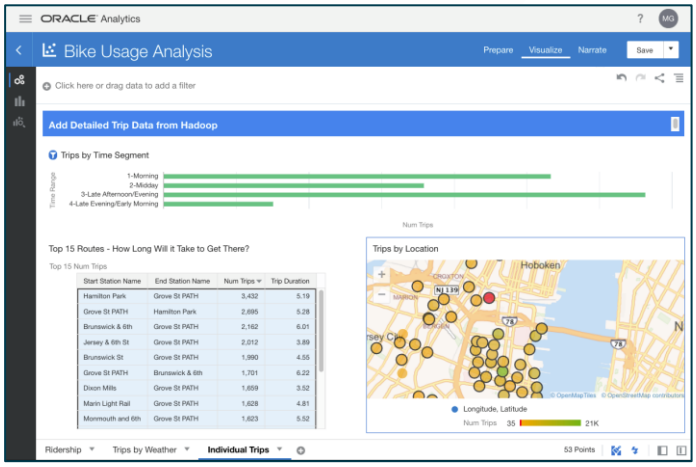

Figure 4. Cloud SQL enables Oracle Analytics Cloud dashboard to show insights from data in Big Data Service, Object Storage and Kafka.

OPEN FOR INNOVATION

Big Data Service embraces the innovations in the big data domain by providing an open environment for innovation, while automated lifecycle management and advanced security ensure organizations do not compromise enterprise-level stability and safety. Organizations are free to deploy other Oracle or third-party software to support new functionality – such as natural language processing and fraud detection – to meet the needs of the application. Support for non-Oracle components is delivered by their respective support channels and not by Oracle.

BIG DATA SERVICE – CUSTOMER BENEFITS

Oracle Big Data Service speeds time to value by providing a cloud-based automated service for big data processing using Cloudera CDH and Apache Spark. Big Data Service provides flexible compute shapes, rich security, and high availability.

- Secure clusters ensure that key data assets remain safe
- Highly Availability ensures your environment remains functional even during infrastructure and software failures.

- Right sized clusters - from small development/POC clusters to comprehensive data lakes
- Workload portability helps maximize your big data investment across on-premise and cloud
- Existing applications are able to gain insights from all of your big data sources using Oracle SQL
- Gain insights and process data using your favorite tools from both Oracle and third parties directly on Big Data Service.

BIG DATA SERVICE INCLUDED SOFTWARE

Software Automatically Installed on Provisioned Instances

- Oracle Linux 7
- Oracle Java – JDK 8
- Cloudera Enterprise (Data Hub Edition) 5.x or 6.x:
  - Cloudera's Distribution including Apache Hadoop (CDH)
  - Cloudera Impala
  - HBase (as well as support for Accumulo)
  - Cloudera Search
  - Apache Spark
  - Apache Kafka
- Cloudera Manager, including:
  - Cloudera Back-up and Disaster Recovery (BDR)
  - Cloudera Navigator

Software License Included but not Installed

- Oracle Big Data Connectors:
  - Oracle SQL Connector for Hadoop
  - Oracle Loader for Hadoop
  - Oracle Machine Learning for Spark
  - Oracle Data Integrator Enterprise Edition
- Oracle Big Data Spatial and Graph

Additional Service (separately priced)

- Oracle Cloud SQL

CONNECT WITH US

Call +1.800.ORACLE1 or visit oracle.com.
Outside North America, find your local office at oracle.com/contact.

blogs.oracle.com/oracle facebook.com/oracle twitter.com/oracle

Integrated Cloud Applications & Platform Services

ORACLE