ORACLE
Cloud Infrastructure
Data Science

Data Science Platforms

Helping IT drive digital transformation

June 2020 | Version 1.01

INTRODUCTION

Digital transformation is an increasingly prevalent buzzword, but what does it mean? George Westerman, principal research scientist with the MIT Sloan Initiative on the Digital Economy, provides an easily digestible definition: “Digital transformation is when companies use technology to radically change the performance or reach of an enterprise”.[1]

As technology advances at a remarkably quick pace, companies must shift and adapt their high-level strategies to keep up. Research attests to the importance of a digital transformation strategy. In a study conducted by marketing research firm Altimeter, 41 percent of the 500 strategists and executives surveyed reported increased market share thanks to digital transformation efforts, 37 percent reported increased customer engagement in digital channels, and 30 percent reported increased customer revenue, among other benefits.[2]

Accurate data is the critical foundation for an effective digital transformation strategy and hiring and expanding data science teams to leverage insights at the enterprise level has become a top priority. As enterprises increasingly scale their data science teams, it falls on IT to support them effectively.

Accurate data is the critical foundation for an effective digital transformation strategy and hiring and expanding data science teams to leverage insights at the enterprise level has become a top priority.

IT managers supporting a large team of data scientists in an enterprise setting are tasked with data governance, as well as providing the infrastructure and tools that data scientists need. The proliferation of data and data science tools and applications available provides opportunities as well as challenges. Data science teams want to use open source applications, and IT teams want to support open source applications. However, provisioning access, providing security audits, and resolving technical issues with open source applications can be a nightmare for IT to manage and secure.

Ultimately, if a company wants to stay ahead of the digital transformation curve, it needs to have a product that supports reproducibility and standardization. A data science platform can help a company meet the demands of its team. This white paper discusses the state of data science today and addresses the problems that a data science platform solves—both from the IT manager perspective, as well as across the larger organization.

THE STATE OF ENTERPRISE DATA SCIENCE

Data science, a discipline that uses data to inform business decisions, is more than just a trendy term. The discipline encompasses many job titles across different industries and organizations, from analytics officer, to actuary, to research scientist. Regardless of the exact title, all of these roles are united in their mission to unlock strategic insights from data to meet business' demand for actionable data. In fact, IBM predicts that demand for data scientists will soar 28 percent by 2020,[3] and IDC predicts that revenue from the sales of big data and analytics tools, applications, and services will increase to more than $187 billion in 2019, up from $122 billion in 2015.[4]

Many companies, such as JetBlue, implement digital transformation strategies with a strong emphasis on data science. The company launched a subsidiary, JetBlue Technology Ventures, to invest in startups that bring machine learning and analytics to the travel market.[5] In March 2016, the company made its first investment in Flyr, a startup that provides predictive analytics and machine learning software to let travelers know when to purchase an airline ticket.[6] Discussing the reasoning behind the investment, President of JetBlue Technology Ventures Bonny Shimi said, "Flyr shares our belief that predictive analytics can provide value to travelers and will change the travel experience in ways we have yet to imagine."

Retail is another industry harnessing the power of big data. "The battle against ecommerce is putting new pressure on brick-and-mortar retailers to fix up stores and deliver a more pleasant experience for shoppers. This means studying data—lots and lots of data," writes Taylor Cromwell of *Bloomberg*.[7] Indeed, a range of retail companies, from Walmart to Nordstrom, are using data-driven insights to implement real-time dynamic pricing and recommendation engines, among other applications.

A data science platform resolves crucial issues surrounding data silos, standardization, resource management, and reproducibility that prevent companies from realizing the full extent of their revenue potential.

The evidence is overwhelming: The marriage of big data and data science is prevalent in every field and sector, and every company wants to leverage data to transform how it does business. However, most aren't. According to Forrester, only 22 percent of companies are actually leveraging big data well enough to get ahead of their competition.[8] What's holding businesses back? In the United States, over a third of companies consider the complexities of IT to be the biggest hurdle to digital transformation.[9] In the next section, we will explore some of the pain points IT teams face as they support increasing numbers of data scientists—and how a data science platform can help.

PAIN POINT #1

Problem: Data Silos

Data silos occur at the enterprise level when multiple teams set up their own data stores based on a use case or for the purpose of isolating data access. IT managers will only grant access to those who truly need it, and also keep databases separated by project, to ensure there won't be resource contention. Even though data silos were built with the best intentions in mind, they often yield more problems than solutions, as they contradict the analytical needs of many organizations where querying across data is required. A data silo, according to the Harvard Business Review, is "a big costly demon" that makes it "prohibitively costly to extract data and put it to other uses."[10] Implementing a data lake, a storage repository that holds a vast amount of data in its native format until it is needed,[11] is the first step towards dealing with the issue of a data silo.

Data silo.

A data store set up for a single use case or to limit data access.

Solution: Data Lakes, Data Warehouses, and Hadoop-Based Systems

To combat the issues surrounding data silos, many enterprises combine, at regular intervals, the data from each separate store into a data lake, which could include a Hadoop Distributed File System (HDFS), S3, or another Hadoop-compatible file system. Another option is to combine the data from each store into a data warehouse such as

Redshift or Vertica. Data lakes and data warehouses are not meant to replace isolated data stores but are rather a solution to combine all data so it can be queried holistically.

Data lake.

A storage repository that holds a vast amount of data in its native format until it is needed.

Data lakes, which have "been well received by enterprises to help capture and store raw data of many types at scale and low cost to perform data management transactions, processing and analytics based on special use cases,"[12] are largely the preferred method of data storage over data warehouses.[13] When an enterprise implements a data lake, they are placing all data into a common location. Hadoop—an open source Java-based programming framework that supports the processing and storage of extremely large data sets—can store the data from a data lake in an HDFS. Once it has been collected into a Hadoop-compatible file system, the data can be queried and combined using a standard set of tools.

In addition, Hadoop distributed file systems address the security issues that data silos aim to solve, as there are a number of technologies able to secure a Hadoop environment. Most commonly, Kerberos is used for user-level authentication in Hadoop-based environments. Additionally, tools such as Apache Sentry, Knox, and Ranger are used to provide more fine-grained authorization to access Hadoop data.

By authenticating through Kerberos, users can deploy Oracle Cloud Infrastructure Data Science—a data science platform—on top of data lakes that are stored in Hadoop distributed file systems, ensuring that data is secure and can be queried holistically across the platform.

PAIN POINT #2

Problem: Lack of Standardization

One way of standardizing data analysis and data science is to bring all of the tools, programming languages, package markers, and software dependencies into one centralized platform. However, due to the size of many enterprises, which can be comprised of hundreds of data scientists tackling a wide breadth of projects, standardization can bring significant challenges and less-than-ideal scenarios, outlined below.

SCENARIO A: LARGE, SHARED REMOTE MACHINES

Data scientists work on remote machines provisioned by IT. IT installs all of the packages needed by data scientists throughout the enterprise. This results in management challenges and difficulty adding new tools, packages, and dependencies as needs diverge across teams.

SCENARIO B: LOW STANDARDIZATION, LACK OF REPRODUCIBILITY

Alternatively, data scientists work on individual machines per team. In these cases, data scientists have additional flexibility, and may be able to configure the tools that they need for their specific task. However, these environments are rarely available across teams and lack oversight from IT.

The use of containerization technologies, such as Docker, capture system dependencies in a lightweight, reproducible way that can be shared across teams.

Solution: Container Technologies

The use of containerization technologies, such as Docker, capture system dependencies in a lightweight, reproducible way that can be shared across teams. IT can implement system configuration in a Docker file and store Docker images in a central repository, enabling individuals to quickly launch the environment they need depending on the project they are working on. Using Oracle Cloud Infrastructure Data Science, IT can set up base environments with the packages, languages, and tools that data scientists need. This gives IT the governance and management over tools and applications, while also empowering data scientists to run self-service analyses.

As open source applications infiltrate the marketplace, it's critical that IT provision settings so that data scientists working across an enterprise have access to the same versions of tools.

PAIN POINT #3

Problem: Lack of Resource Management

All data science projects require compute resources, whether you're using a local laptop or a powerful cluster of cloud-based servers. Often in an enterprise, many large servers are shared across a team of data scientists. From an IT administration standpoint, provisioning and managing these servers can be a time-consuming endeavor. Because data science projects vary so much in resource requirements, a single user could easily be consuming all of a machine's compute power, leaving it unavailable for other users.

Provisioning and managing compute resources can be a time-consuming endeavor.

To avoid the misallocation of resources, the IT team should have complete control over the cluster, and data scientists should only be able to choose from a pre-determined bucket of compute sizes. Additionally, IT should be able to scale out the available resources as needed without downtime, a problem that is currently costing businesses US$700 billion a year.[14]

Solution: Using Docker Swarm in the DataScience.com Platform

Oracle Cloud Infrastructure Data Science leverages Docker Swarm for container orchestration. This allows IT to provision a pool of servers and configure the sizing options from which a data scientist can choose.

Oracle Cloud Infrastructure Data Science allows for on-demand servers for customers running on public clouds such as Amazon Web Services (AWS). The platform provides resource management tools for IT to show which users are using resources within particular projects. It also gives IT the governance to shut down a user's container if it is left running too long or is consuming too many resources.

Without a standardized infrastructure and workflow—as well as a means of configuring, centralizing, discovering, and deploying these components—enterprises can't achieve reproducibility.

PAIN POINT #4

Problem: Low Reproducibility

Imagine the scenario where a data scientist leaves a company unexpectedly, and there is no documentation available regarding the best practices surrounding his or her work. When a new data scientist joins the team and is assigned that same project, he or she will have no idea what to do and will likely lean heavily on IT to get up to speed.
For many IT teams supporting a large team of data scientists, this scenario is all too realistic. The underlying pain points. A lack of reproducibility. As Ian Swanson, Oracle VP of Product Management, Machine Learning, explains: "A critical step to gaining reproducibility in data science is defining all of the different steps involved, from ingesting the first piece of data to deploying a model in production. A data science platform will help to create and manage the workflows involving these steps."
Without a standardized infrastructure and workflow—as well as a means of configuring, centralizing, discovering, and deploying these components—enterprises can't achieve reproducibility.

Solution: An Enterprise Data Science Platform

Enterprise data science platforms, such as Oracle Cloud Infrastructure Data Science, effectively tackle the issue of lack of reproducibility by:

- Centralizing all of the assets—tools, code, data—needed to do data science across the enterprise.
- Providing version control functionality so that data scientists can track versions of code, as well as deploy multiple versions and test them against each other.

- Showcasing work around the enterprise and making it accessible with features such as projects, centralized outputs, and search. IT has governance over this functionality by implementing role-based access control.
- Providing standardized mechanisms for promoting work to production. This includes running scripts as scheduled jobs, or providing microservices, a way of breaking up a huge job into smaller parts that can be easily maintained. In the case of Oracle Cloud Infrastructure Data Science, this means deploying a predictive model or code as an API, which eliminates the burden on IT and engineering on a per-project basis. Otherwise, there would be a number of initiatives involved before deploying to production, including refactoring code, rewriting into a production stack language, and testing performance, among other steps.[15]

CONCLUSION: SCALING SUCCESSFULLY

As IT managers continue to support larger teams of data scientists within organizations, data science platforms will continue to rise in prominence as an effective and necessary means of scaling a digital transformation strategy. “An enterprise data platform helps data scientists get the most value out of their data by managing the data analytics lifecycle and standardizing routine processes while enforcing security and governance,” said Ian Swanson, VP of Product Management, Machine Learning, at Oracle.[16] From the perspective of an IT manager, a data science platform also resolves crucial issues surrounding data silos, standardization, resource management, and reproducibility that currently prevent many companies from realizing the full extent of their revenue potential.

> “An enterprise data platform helps data scientists get the most value out of their data by managing the data analytics lifecycle and standardizing routine processes while enforcing security and governance.”
>
> **Ian Swanson**
>
> VP, Product Management, Machine Learning, Oracle

REFERENCES

1. CIO, “What is digital transformation? A necessary disruption,” July 2017.
2. Altimeter, “The 2016 State of Digital Transformation,” 2016.
3. Forbes, “IBM Predicts Demand for Data Scientists Will Soar 28% by 2020,” May 2017.
4. InformationWeek, “Big Data, Analytics Sales Will Reach $187 Billion by 2019,” June 2016.
5. CIO, “An Inside Look at 10 real-world digital transformation success stories,” June 2017.
6. CIO, "JetBlue CIO pilots VC arm in search of revenue growth," May 2016.
7. Bloomberg, “Here’s a Retail Job That’s Still in High Demand: Data Scientist,” August 2017.
8. Forrester, “Data Science Platforms Help Companies Turn Data Into Business Value,” December 2016.
9. Dynatrace, “The Global Digital Performance & Transformation Audit,” 2017.
10. Harvard Business Review, “Breaking Down Data Silos,” December 2016.
11. TechTarget, "Definition: Data Lake"
12. 1.2 INFORMS Analytics, “DATA LAKES: The Biggest Big Data Challenges,” November/December 2016.
13. KDNuggets, “Data Lake vs. Data Warehouse: Key Differences,” September 2015.
14. IHS Markit, “Businesses Losing $700 Billion a Year Due to Downtime, Says IHS,” January 2016.
15. DataScience.com, Navigating the Pitfalls of Model Deployment, December 2016.
16. Datanami, “Data Science Platforms Seen as Decision-Makers,” January 2017.

CONNECT WITH US

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

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