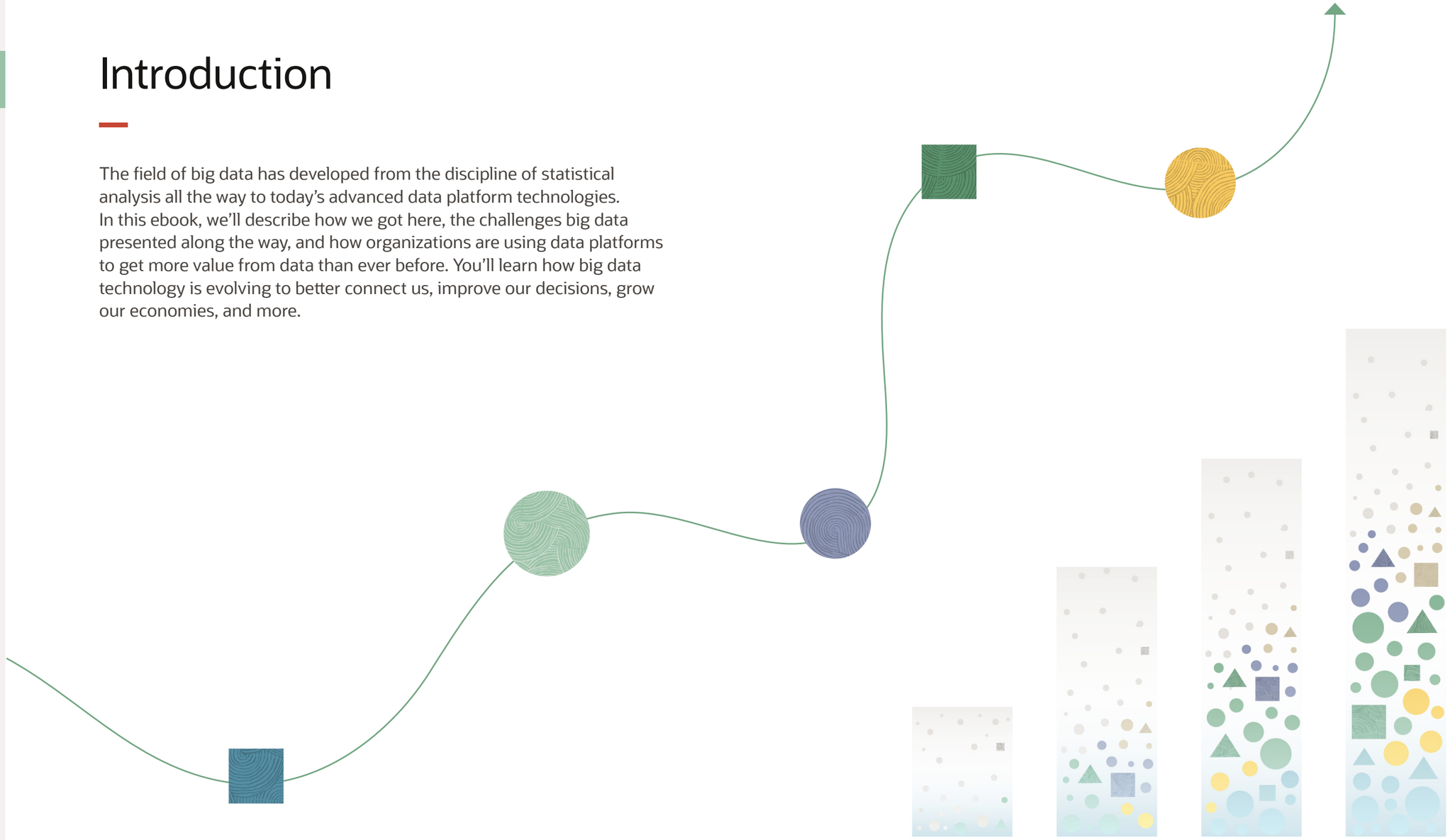
ORACLE CLOUD Infrastructure

The Evolution of Big Data and the Future of the Data Platform

How organizations use data platforms to get more value from data

Introduction

The field of big data has developed from the discipline of statistical analysis all the way to today's advanced data platform technologies. In this ebook, we'll describe how we got here, the challenges big data presented along the way, and how organizations are using data platforms to get more value from data than ever before. You'll learn how big data technology is evolving to better connect us, improve our decisions, grow our economies, and more.



Big data beginnings

Put simply, big data is a concept describing data sets that exceed the size that can be managed by traditional tools. It's defined by three Vs: variety, volume, and velocity. The growing variety of data sources that arrives in increasing volumes and with more velocity (the high rate at which data is received and acted on).

The roots of big data come from "business intelligence," a term IBM (PDF) coined in 1958, defining it as

> "the ability to apprehend the interrelationships of presented facts in such a way as to guide action towards a desired goal."
>
> IBM, 1958

The 1960s and '70s saw significant advancements in data technology with the development of mainframes and databases. The 1980s saw the emergence of personal computers and client-server computing and, along with that, relational databases and SQL (Structured Query Language). With each of these breakthroughs, the utility and volume of data grew.

Data volumes exploded in the '90s with the rise of the internet, ecommerce, and search technologies. For transactional databases, this meant new architectures to support more performance, scalability, and redundancy. At the same time, the need for business intelligence across these data volumes drove companies to create new types of databases—data warehouses, specialized relational databases optimized for analytics—to store curated data from a wide variety of sources. The data warehouses became core infrastructure that companies used to track their operations, complete reporting, perform analysis, and support decision-making.

New big data approaches

Around 2005, we entered the era of Web 2.0, when companies began to realize just how much data users generated through social media and other online services. Data of all types, structured and unstructured, needed to be collected, processed, and analyzed. Current technologies couldn't process it, at least not economically. A new approach was needed.

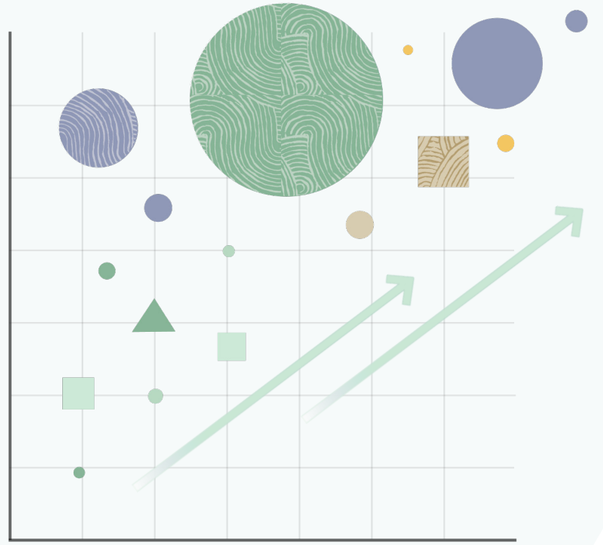
Google published a paper on MapReduce, a programming model that defined a system for processing large data sets. Yahoo got involved in the project, and Hadoop was created. In 2008, Yahoo released Hadoop to the Apache Software Foundation, followed by the Apache Software Foundation releasing Apache Hadoop 1.0 in 2011.

Hadoop, an open source framework, accelerated the utility and growth of big data. The Hadoop Distributed File System is a storage system that can distribute data across clusters of computers. MapReduce enables parallel processing of that distributed data to increase performance. The combination enabled the big data use cases that accelerated the digital economy, such as the 360-degree views of ecommerce customers. These use cases had previously been impossible or cost prohibitive to achieve.

The Hadoop framework rapidly expanded with tools for deploying and managing clusters, scheduling processes, querying data, and more. Spark, an open source data processing engine for large data sets, became popular because it enabled computational speed, scalability, and programmability for big data—specifically with applications for streaming data, graph data, machine learning (ML), and artificial intelligence (AI). Spark stores and processes data in memory. This is key to Spark's performance because it lets applications avoid slow disk accesses.

Big data accelerated the digital economy,

enabling use cases that had previously been impossible or cost-prohibitive to achieve.



Big data challenges

Many companies that adopted these big data technologies have found that managing the various open source Hadoop tools was complex and expensive. Companies, such as Cloudera and Hortonworks, sold packaged versions of the open source projects to address that market.

Even with these packaged solutions, a major challenge for companies adopting Hadoop is managing the required data center infrastructure. Clusters must be able to scale to hundreds or even thousands of nodes for certain jobs. They must be available when needed but may be idle for periods as well. Those requirements make it difficult to find the right balance between economics and the availability of the cluster infrastructure.

The cost of storage also becomes an issue as organizations struggle to keep pace with the amount of data they need to store. Tiering of storage to support production, backups, archives, and so forth is crucial to contain costs.

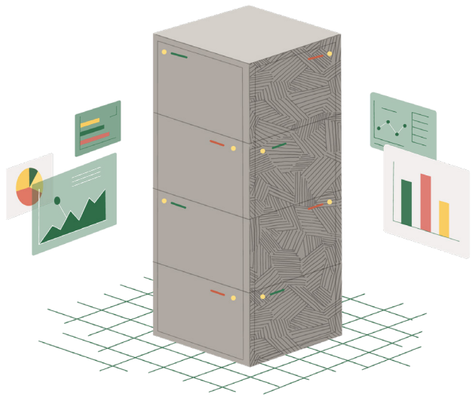
The emergence of public clouds in 2010 promised to address these challenges. With flexible compute and reliable low-cost storage, they became an attractive option for building one's own data clusters.

Data lakes

The emergence of public clouds had a profound impact on the way organizations could tackle big data challenges. The availability of cheap, reliable, and infinitely scalable storage let companies ingest and store the data raw and unchanged, instead of cleaning, transforming, and aggregating it before storage. That, in turn, enabled new methods of analyzing the data that previously weren't available.

James Dixon, then chief technology officer at Pentaho, coined the term "data lake" for this new approach. Rather than creating isolated data warehouses, a data lake promised to be a single repository for all of a company's information.

Data lakes can be built with Hadoop technologies or with object storage and managed data services provided by a cloud provider. By delegating the infrastructure work and applications management to a cloud provider, companies can decrease the IT work of big data tasks and focus on data management.



The following are some of the data tools that many cloud providers offer their users:

Object Storage

Enables organizations to store any type of data in its native format—this is ideal for building modern applications that require scale and flexibility

Data Integration

Easy-to-use tools that connect to public and private data sources such as databases and applications and reliably transfer and synchronize the data to the datastores in the data lake

Data Preparation

Visual tools to create data transformations between the source and the target

Data catalog

An inventory of enterprisewide data assets to help search, explore, and govern data in the data lake

Data streaming

Lets organizations process data in real time, enabling resilient stream processing operations such as filters, joins, maps, aggregations, and other transformations

Data management

Hadoop, Spark, databases, and query tools that help organizations manage data across all stores in the data lake

Analytics

Tools to help organizations understand and discover trends in their data and use them to guide decision-making

Using those tools, companies can start data lakes for their unstructured data on a small scale and continually expand them with new data types, data sources, and applications to derive value from the data.

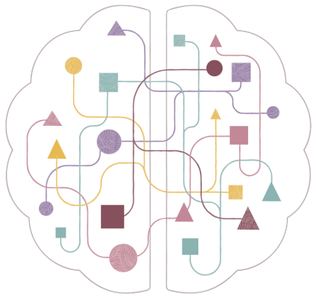
Learn how to build a data lake →

Data platforms

Remember transactional databases and data warehouses? They continue to be core building blocks for managing structured data for most companies. But now data lakes with unstructured data are also becoming critical as more companies leverage their unstructured data for business insights. Increasingly, companies are looking at databases, data warehouses, and data lakes as equal components of their entire data estate. Integrated management across these silos is required for comprehensive management and analytics across all of a company's data.

That's what a data platform offers. A data platform combines the best features of transactional databases and data warehouses with the best features of data lakes. A data platform will reduce data redundancy, eliminate the costs of maintaining multiple data storage systems, support a wide variety of workloads, and improve data security.

A data platform's modern, open architecture can store and process all of an organization's data, including structured and unstructured data, while enabling users to access information more quickly and start putting it to work. Organizations use data platforms to capture, manage, and analyze data in real time to improve customer experience, reduce fraud, and speed up time to market.



Data scientists spend

39% of their time

curating and preparing data before it can be used.

AI and ML

AI and ML are the next disruptors in big data technology. With AI and ML, computers can recognize the content of images, transcribe spoken language, read texts, and understand the sentiment of social media responses. Where data platforms were at first tools to simply collect data of all types, with AI and ML they can now understand the stored data and use that to initiate actions or support decisions.

The integration of AI and ML into data platforms has enabled many new use cases that were never achievable previously.

Business use cases

There are hundreds of ways big data can give businesses a competitive advantage. Here, we'll explore just a few examples of how industries have used big data to go beyond measuring and counting to being able to predict and understand.

Financial services

Whether it's capturing new market opportunities or reducing fraud, financial services organizations have converted big data into a competitive advantage. With big data, financial institutions can

- Identify patterns that indicate fraud and streamline regulatory reporting
- Gain a better understanding of market trends and customer needs, which can improve decision-making about new products and services
- Detect potential fraud patterns and adhere to regulations

Healthcare

Hospitals, healthcare companies, and researchers produce massive amounts of data. With big data, healthcare professionals can

- Identify disease genes and biomarkers to help patients pinpoint health issues they may face in the future
- Provide better treatment and improve the quality of care without increasing costs
- Detect potential insurance fraud by flagging certain behaviors for further examination

Manufacturing

The digital revolution has empowered manufacturers to find ways to harness all the data they generate. With big data, manufacturers can

- Predict equipment failures
- Assess production processes, proactively respond to customer feedback, and anticipate future demands
- Better understand the flow of production lines and determine the cause of delays

Retail

Big data is used across all stages of the retail process. With big data, retailers can

- Predict customer demand and launch new products
- Identify a company's most loyal customers and target them with special offers
- Use predictive technology to keep shelves stocked and avoid supply chain disruptions

Telecommunications

The popularity of smartphones and other mobile devices has given telecommunications companies tremendous growth opportunities and an ever-expanding volume of data. Big data lets telecommunications companies

- Identify areas with excess capacity and reroute bandwidth as needed
- Predict overall customer satisfaction by analyzing the data they already have about service quality and convenience
- Improve understanding of customer behavior to design new products and features

Download the Big Data Use Cases Ebook

Conclusion

Big data is less about individual pieces of data and more about gleaning value and meaning from it. Today, data is coming from more sources, in more formats, more quickly than ever before. With the right data platform in place, organizations can get value from their data so they can make faster, data-driven decisions.

Let Oracle help your business evolve

Simplify big data application delivery with Apache Spark →

Make it easier to build managed data lakes →

Enable self-service data discovery and governance →

ORACLE CLOUD
Infrastructure