What is Big Data?

Brady Pinter

Belmont University

February 3, 2026

Outline

What is Big Data?

The Three Phases

Files and Storage

The Future

Conclusion

Resources

Thinking Big

- According to the IDC, there was "33 zettabytes of data in 2018 and a predicted amount of 175 zettabytes in 2025."
- If you were to store 175 zettabytes of data as CDs, it would circle the Earth 222 times.
- And to give a reference, a zettabyte is 1 trillion gigabytes.

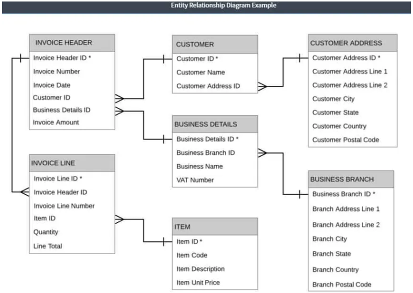
What is Big Data

- Google has defined Big Data as ”extremely large and diverse collections of structured, unstructured, and semi-structured data that continue to grow exponentially over time.”
- This large amount of information has been seen as ”the new oil” to fuel business growth and innovation as we move into the foreseeable future.

What is Big Data (Cont.)

- The idea of Big Data was coined by John R. Mashey in the 1990s as a way to describe the expansive explosion of collected information in that time.
- Big Data has evolved through multiple phases which each faced distinct problems which needed to be solved.
- The three big phases of Big Data can be split down based on the software each phase had.
  - RDBMS
  - Google File System, MapReduce, BigTable, and Hadoop!
  - Apache Spark

What is a RDBMS?



Entity Relationship Diagram Example

INVOICE HEADER
- Invoice Header ID *
- Invoice Number
- Invoice Date
- Customer ID
- Business Details ID
- Invoice Amount

CUSTOMER
- Customer ID *
- Customer Name
- Customer Address ID

CUSTOMER ADDRESS
- Customer Address ID *
- Customer Address Line 1
- Customer Address Line 2
- Customer City
- Customer State
- Customer Country
- Customer Postal Code

BUSINESS DETAILS
- Business Details ID *
- Business Branch ID
- Business Name
- VAT Number

INVOICE LINE
- Invoice Line ID *
- Invoice Header ID
- Invoice Line Number
- Item ID
- Quantity
- Line Total

ITEM
- Item ID *
- Item Code
- Item Description
- Item Unit Price

BUSINESS BRANCH
- Business Branch ID *
- Branch Address Line 1
- Branch Address Line 2
- Branch City
- Branch State
- Branch Country
- Branch Postal Code

Figure 1: An example of a RDBMS

- A Relational Database Management System (RDBMS) is a system that allows someone to create, manage, interact, and administer data that is held in relation tables.
- It works well with data that is organized and structured well, and its strict syntax works well to keep data integrity strong across all tables.

Downsides of RDBMS

- Even with all the pros that come with using a RDBMS, it still has many downsides that lead companies like Google to invent their own systems.
- Some of the major issues with RDBMS included:
  - It struggles to handle unstructured data. (This includes data like emails, images, or data with odd-syntax)
  - Its limited scalability makes it tough to use when query large amounts of information.

Building Beyond : Google File System, MapReduce, and Bigtable

- Due to the weaknesses of RDBMS, Google needed to invent their own systems to build and search the Internet's documents. This resulted in the creation of the Google File System, MapReduce, and Bigtable.
- Each software supplied the necessary tools to accomplish the tasks Google was looking to complete.

Google File System

- The Google File System provided a fault-tolerate and distributed file system which could be held across multiple servers in a cluster farm.
- This offered a scalable storage solution for both structured and unstructured data.

MapReduce

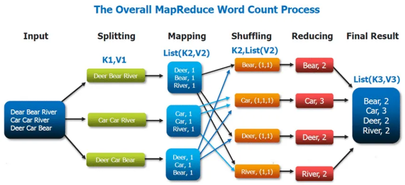


Figure 2: An example of MapReduce

- MapReduce introduced new parallel programming paradigms based on functional programming solutions of the time.
- This facilitated the large-scale processing of data distributed across the Google File System and held in BigTables.

BigTable

- Using data that is held in the Google File System, BigTable provides the organization of that data into a scalable database structure.
- This organization is then queried using MapReduce to process the data and send it to a respective output location.

Hadoop and Innovation

- Even with the software google created, more innovation was needed to streamline queries and data management into one software.
- The invention of Hadoop and Yahoo! allowed for the creation of the Hadoop File System (HDFS) which included MapReduce as a framework for distributed computing.

Limitations of Hadoop

Even with everything the HDFS introduced, it did have its own personal shortcomings.

- It was cumbersome to handle, and its operational complexity gave it a steep learning curve that many struggled to grasp and understand well.
- Using MapReduce required a lot of setup and boilerplate code, which also had brittle fault tolerance.
- During large queries, the intermediate writing of results to disk slowed down operations, and caused many large jobs to take hours or days to run.

Apache Spark is Created

Apache Spark can be described as "a unified engine designed for large-scale distributed data processing on premises in data centers or in the cloud"

- Apache Spark provides in-memory storage for intermediate computations, making it much faster than the HDFS.
- It also incorporates many libraries as well, making it usable in machine learning processes, and interaction queries using SQL.

How Apache Spark Works

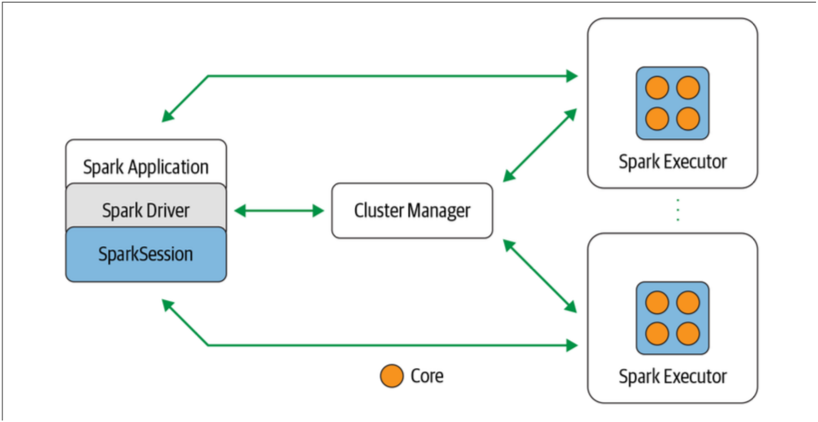


Figure 3: An example of a Spark Cluster

Apache Spark's Four Characteristics

The design philosophy of Apache Spark revolved around four key characteristics:

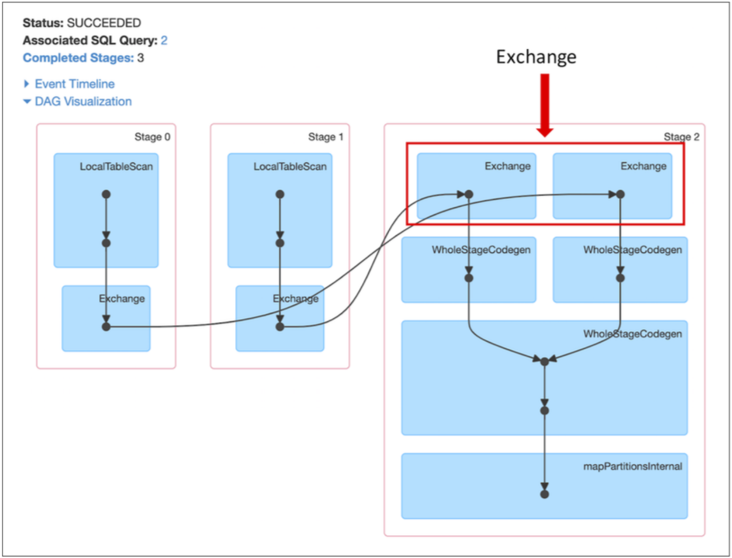
- Speed - It was one of the main goals of Spark to be fast. It has done this by building its queries as directed acyclic graphs (DAGs). This formatting allows tasks to be "decomposed into tasks that are then executed in parallel across workers on the (computer) cluster".
- Ease of Use - Spark works to be easy to understand. Its simple logical data structure is built upon a system called the Resilient Distributed Dataset (RDD). This system allows you to program big data applications using familiar languages.

Apache Spark's Four Characteristics (Cont.)

The design philosophy of Apache Spark revolved around four key characteristics:

- Modularity - The ease of use is connected to the modularity. Spark operations can be applied across a wide variety of supported programming languages like Scala, Java, Python, SQL, and R.
- Extensibility - Due to the fact that Apache Spark separates storage from computing, you can use large Apache Spark ecosystem to hold your data and focus on computing with the Spark framework only.

Apache DAGs at Work



Status: SUCCEEDED

Associated SQL Query: 2

Completed Stages: 3

Figure 4: A example of a Spark DAG

Apache Spark Ecosystem

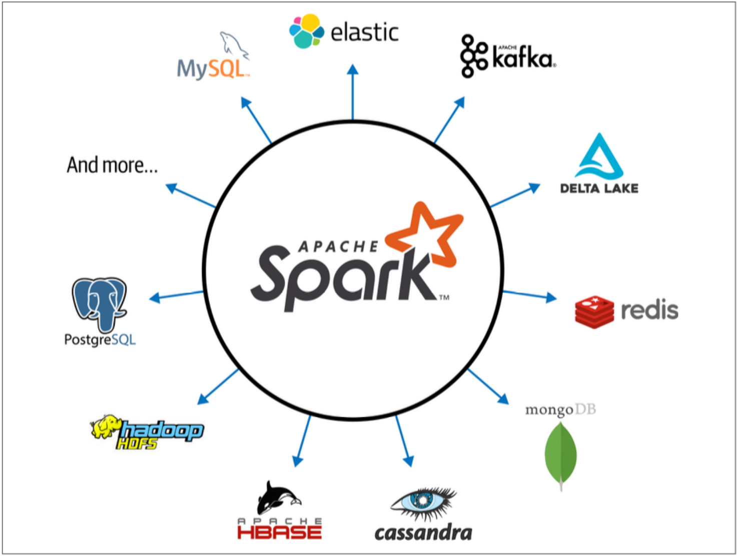


Figure 5: The Apache Spark Ecosystem

Parquet Files

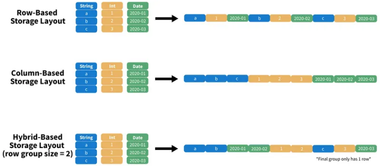

Row-Based Storage Layout
String
a
b
c
Int
1
2
3
Date
2020-01
2020-02
2020-03
a 1 2020-01 b 2 2020-02 c 3 2020-03
Column-Based Storage Layout
String
a
b
c
Int
1
2
3
Date
2020-01
2020-02
2020-03
a b c 1 2 3 2020-01 2020-02 2020-03
Hybrid-Based Storage Layout (row group size = 2)
String
a
b
c
Int
1
2
3
Date
2020-01
2020-02
2020-03
a b 2020-01 2020-02 1 2 c 3 2020-03
Final group only has 1 row

Figure 6: An example of column vs. row based orientation.

- A parquet file is a column-oriented data storage format which is apart of the Apache Ecosystem.
- Parquet files are known for its data compression and encoding schemes which can handle large amounts of information in bulk.

The Future of Big Data

The future of Big Data will be defined by cloud computing, and a variety of emerging technologies which hope to shape how we use data to assist business and everyday decisions.

- The surge of AI usage across all fields of computing will lead to its integration in a variety of software uses to query big data. It is clear that AI will be used to optimize our data and allow us to reach answers faster.
- The emergence of technology like edge computing will allow us to analyze data closer to its source, and limit the amount that is sent to a centralized system for processing.

Conclusion

The development and use of Big Data in business solutions will see rapid growth in the coming future due to the amount of data that is on the internet.

- The development and use of software like Apache Spark will give us the tools we need to query, analyze, and develop solutions to complex data problems for the foreseeable future.
- If you want to learn more, I can send you a PDF of a great textbook resource for using spark.

Resources

- A lot of the information and research for this presentation was provided by the Learning Spark 2.0 textbook. The introduction section of that book gave me great knowledge on the history of Big Data, and how Apache Spark works.