Intro to Hadoop and Big Data

Revolutionizing Data Processing at Scale

Vinod Kumar Kayartaya (vinod@vinod.co)

What is Big Data?

- Definition:
  - Big Data refers to the massive volume of data that is too large, complex, and fast-growing to be processed and analyzed using traditional data processing tools.

Basic Units

1. **Bit (b)**: The smallest unit of data, representing a binary value (0 or 1).
2. **Byte (B)**: 8 bits.

Commonly Used Units

3. **Kilobyte (KB)**: 1,024 bytes (2^{10} bytes).
4. **Megabyte (MB)**: 1,024 KB or 1,048,576 bytes (2^{20} bytes).
5. **Gigabyte (GB)**: 1,024 MB or 1,073,741,824 bytes (2^{30} bytes).
6. **Terabyte (TB)**: 1,024 GB or 1,099,511,627,776 bytes (2^{40} bytes).
7. **Petabyte (PB)**: 1,024 TB or 1,125,899,906,842,624 bytes (2^{50} bytes).
8. **Exabyte (EB)**: 1,024 PB or 1,152,921,504,606,846,976 bytes (2^{60} bytes).
9. **Zettabyte (ZB)**: 1,024 EB or 1,180,591,620,717,411,303,424 bytes (2^{70} bytes).
10. **Yottabyte (YB)**: 1,024 ZB or 1,208,925,819,614,629,174,706,176 bytes (2^{80} bytes).

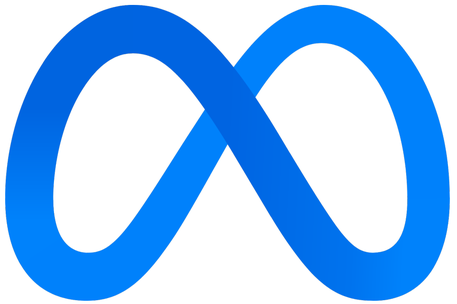
Larger Units (Rarely Used)

11. **Brontobyte**: 1,024 YB or 2^{90} bytes.
12. **Geopbyte**: 1,024 Brontobytes or 2^{100} bytes.

1. Google

- **Data Volume:** Over 20 petabytes (PB) per day.
- **Details:** Google handles over 3.5 billion searches per day, processes large-scale datasets from various services like YouTube, Google Maps, and Gmail, and analyzes massive amounts of data to optimize search results and ad targeting.

2. Facebook (Meta)



- **Data Volume:** Over 4 petabytes (PB) of data per day.
- **Details:** With billions of users sharing photos, videos, and text, Facebook processes and stores massive volumes of data daily. This includes data from platforms like Instagram and WhatsApp, as well as metadata from interactions and ad engagements.

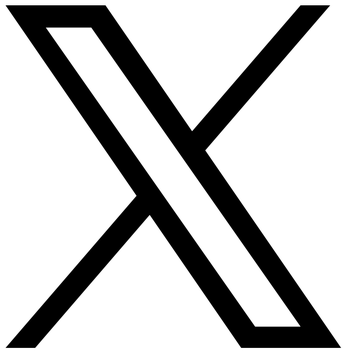
3. Amazon

- **Data Volume:** Hundreds of terabytes (TB) per day.
- **Details:** Amazon generates vast amounts of data from its e-commerce platform, AWS (Amazon Web Services), and other services like Alexa. This data includes customer purchase history, website traffic, and cloud service usage.

4. Netflix

- **Data Volume:** Over 100 terabytes (TB) of data per day.
- **Details:** Netflix processes large volumes of data related to user viewing habits, streaming quality, and recommendations. The data helps Netflix refine its recommendation algorithms and improve user experience.

5. Twitter



- **Data Volume:** Around 500 million tweets per day, resulting in several terabytes (TB) of data daily.
- **Details:** Twitter generates and processes data from tweets, retweets, likes, user interactions, and trending topics, all of which contribute to its vast data ecosystem.

6. Airbnb

- **Data Volume:** Several terabytes (TB) of data per day.
- **Details:** Airbnb collects data from bookings, user reviews, property listings, and user interactions. This data is used to optimize search algorithms, pricing models, and customer experiences.

What is Big Data?

- Characteristics:
  - Volume: Huge amounts of data generated every second.
  - Velocity: The speed at which data is generated and processed.
  - Variety: Different types of data—structured, unstructured, and semi-structured.
  - Veracity: Quality and accuracy of the data.
  - Value: The potential insights and benefits derived from analyzing Big Data.

Challenges of Big Data

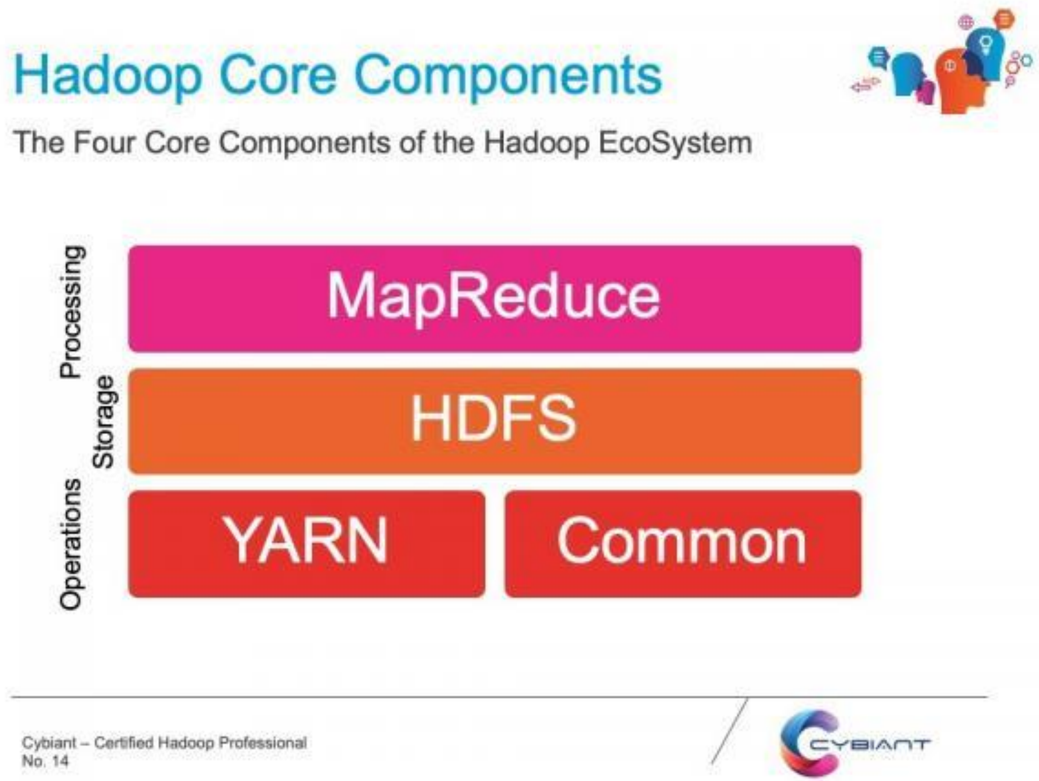
- Data Storage: Traditional databases struggle to store and manage massive datasets.
- Data Processing: Complexity in processing and analyzing large datasets quickly.
- Scalability: Difficulty in scaling infrastructure to accommodate growing data needs.
- Data Integration: Combining different types of data from various sources.

Introduction to Hadoop

- What is Hadoop?
  - An **open-source** framework designed to store and process **large datasets** across **distributed computing environments**.

Introduction to Hadoop

- Key Components:
  - Hadoop Distributed File System (HDFS): Stores large data files across multiple machines.
  - MapReduce: A programming model for processing large datasets in parallel across a Hadoop cluster.
  - YARN (Yet Another Resource Negotiator): Manages and schedules resources across the cluster.
  - Hadoop Common: Provides common utilities and libraries required by other Hadoop components.

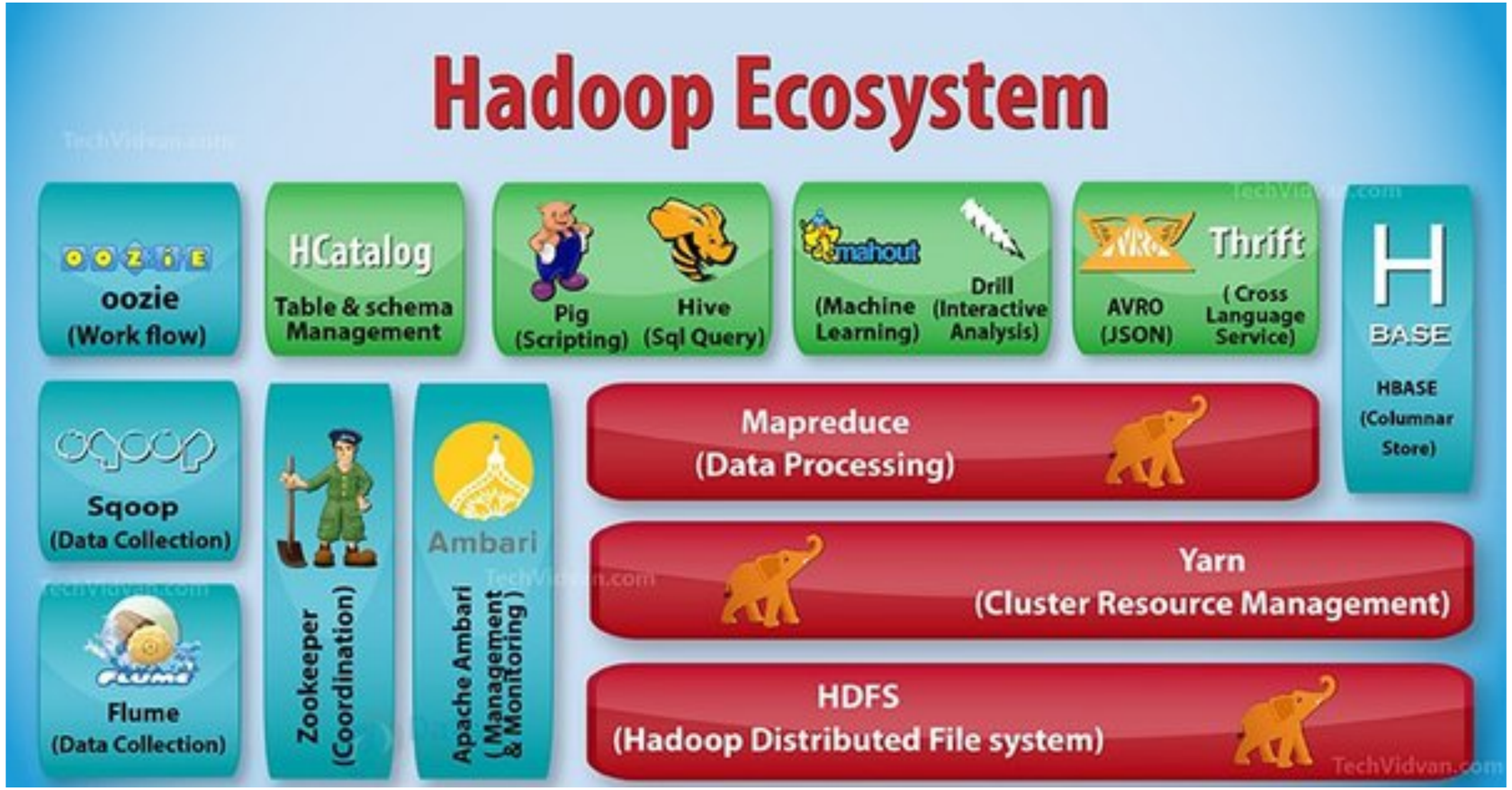


Why Hadoop for Big Data?

- Scalability: Easily scales from a single server to thousands of machines.
- Fault Tolerance: Automatically handles hardware failures by replicating data across multiple nodes.
- Cost-Effectiveness: Utilizes commodity hardware, making it more affordable for large-scale data processing.
- Flexibility: Can handle all types of data—structured, unstructured, and semi-structured.

Hadoop Ecosystem

The Hadoop ecosystem includes various tools and technologies for managing and analyzing Big Data.



Hadoop Ecosystem

- Key Tools:
  - Hive: Data warehousing and SQL-like query language.
  - Pig: High-level data flow language for processing large datasets.
  - HBase: NoSQL database that runs on top of HDFS.
  - Spark: Fast and general-purpose cluster-computing system.
  - Oozie: Workflow scheduler for managing Hadoop jobs.

Real-World Applications of Hadoop

- Industries: Finance, Healthcare, Retail, Telecom, and more.
- Use Cases: Fraud detection, customer behavior analysis, recommendation systems, and large-scale data processing.

https://cwiki.apache.org/confluence/display/hadoop2/PoweredBy#PoweredBy-PoweredbyApacheHadoop

- *A9.com - Amazon**
  - *We build Amazon's product search indices using the streaming API and pre-existing C++, Perl, and Python tools.*
  - *We process millions of sessions daily for analytics, using both the Java and streaming APIs.*
  - *Our clusters vary from 1 to 100 nodes*
- *Accela Communications*
  - *We use an Apache Hadoop cluster to rollup registration and view data each night.*
  - *Our cluster has 10 1U servers, with 4 cores, 4GB ram and 3 drives*
  - *Each night, we run 112 Hadoop jobs*
  - *It is roughly 4X faster to export the transaction tables from each of our reporting databases, transfer the data to the cluster, perform the rollups, then import back into the databases than to perform the same rollups in the database.*
- *Adobe*
  - *We use Apache Hadoop and Apache HBase in several areas from social services to structured data storage and processing for internal use.*
  - *We currently have about 30 nodes running HDFS, Hadoop and HBase in clusters ranging from 5 to 14 nodes on both production and development. We plan a deployment on an 80 nodes cluster.*
  - *We constantly write data to Apache HBase and run MapReduce jobs to process then store it back to Apache HBase or external systems.*
  - *Our production cluster has been running since Oct 2008.*
- *adyard*
  - *We use Apache Flume, Apache Hadoop and PApache ig for log storage and report generation as well as ad-Targeting.*
  - *We currently have 12 nodes running HDFS and Pig and plan to add more from time to time.*
  - *50% of our recommender system is pure Pig because of it's ease of use.*
  - *Some of our more deeply-integrated tasks are using the streaming API and ruby as well as the excellent Wukong-Library.*
- *Able Grape - Vertical search engine for trustworthy wine information*
  - *We have one of the world's smaller Hadoop clusters (2 nodes @ 8 CPUs/node)*
  - *Hadoop and Apache Nutch used to analyze and index textual information*
- *Adknowledge - Ad network*
  - *Hadoop used to build the recommender system for behavioral targeting, plus other clickstream analytics*
  - *We handle 500MM clickstream events per day*
  - *Our clusters vary from 50 to 200 nodes, mostly on EC2.*
  - *Investigating use of R clusters atop Hadoop for statistical analysis and modeling at scale.*
- *Aguja- E-Commerce Data analysis*
  - *We use hadoop, pig and hbase to analyze search log, product view data, and analyze all of our logs*
  - *3 node cluster with 48 cores in total, 4GB RAM and 1 TB storage each.*
- *Alibaba*
  - *A 15-node cluster dedicated to processing sorts of business data dumped out of database and joining them together. These data will then be fed into iSearch, our vertical search engine.*
  - *Each node has 8 cores, 16G RAM and 1.4T storage.*
- *AOL*
  - *We use Apache Hadoop for variety of things ranging from ETL style processing and statistics generation to running advanced algorithms for doing behavioral analysis and targeting.*
  - *The cluster that we use for mainly behavioral analysis and targeting has 150 machines, Intel Xeon, dual processors, dual core, each with 16GB Ram and 800 GB hard-disk.*

Conclusion

- Key Takeaways:
  - Hadoop is essential for processing and analyzing Big Data.
  - Its scalability, fault tolerance, and cost-effectiveness make it the preferred choice for organizations handling vast amounts of data.
  - The Hadoop ecosystem provides a comprehensive suite of tools for a wide range of Big Data applications.