BIG DATA

Prepared By:-
EduTechLearners

How much time did it take?

- Excel : Have you ever tried a pivot table on 500 MB file?
- SAS/R : Have you ever tried a frequency table on 2 GB file?
- Access: Have you ever tried running a query on 10 GB file
- SQL: Have you ever tried running a query on 50 GB file

Can you think of ?

- Can you think of running a query on 20,980,000 GB file.
  - What if we get a new data set like this, every day?
  - What if we need to execute complex queries on this data set everyday ?
  - Does anybody really deal with this type of data set?
  - Is it possible to store and analyze this data?
- Yes Google deals with more than 20 PB data everyday

In fact, in a minute

- **Email** users send more than 204 million messages;
- **Mobile Web** receives 217 new users;
- **Google** receives over 2 million search queries;
- **YouTube** users upload 48 hours of new video;
- **Facebook** users share 684,000 bits of content;
- **Twitter** users send more than 100,000 tweets;
- **Consumers** spend $272,000 on Web shopping;
- **Apple** receives around 47,000 application downloads;
- **Brands** receive more than 34,000 Facebook 'likes';
- **Tumblr** blog owners publish 27,000 new posts;
- **Instagram** users share 3,600 new photos;
- **Flickr** users, on the other hand, add 3,125 new photos;
- **Foursquare** users perform 2,000 check-ins;
- **WordPress** users publish close to 350 new blog posts.

And this is one year back.. Damn!!

What is BIG DATA?

- Collection of data sets so large and **complex** that it becomes **difficult to process** using on-hand database management tools or traditional data processing applications
- "Big Data" is the data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it*
- 'Big Data' is similar to 'small data', but bigger in size
- An aim to solve new problems or old problems in a better way
- Big Data generates value from the storage and processing of very large quantities of digital information that cannot be analyzed with traditional computing techniques.

Three Characteristics of Big Data V3s

Volume

- Data quantity

Velocity

- Data Speed

Variety

- Data Types

1st Character of Big Data Volume

•A typical PC might have had 10 gigabytes of storage in 2000.

•Today, Face book ingests 500 terabytes of new data every day.

•Boeing 737 will generate 240 terabytes of flight data during a single flight across the US.

• The smart phones, the data they create and consume; sensors embedded into everyday objects will soon result in billions of new, constantly-updated data feeds containing environmental, location, and other information, including video.

2nd Character of Big Data Velocity

- Click streams and ad impressions capture user behavior at millions of events per second
- high-frequency stock trading algorithms reflect market changes within microseconds
- machine to machine processes exchange data between billions of devices
- infrastructure and sensors generate massive log data in real-time
- on-line gaming systems support millions of concurrent users, each producing multiple inputs per second.

3rd Character of Big Data Variety

- Big Data isn't just numbers, dates, and strings. Big Data is also geospatial data, 3D data, audio and video, and unstructured text, including log files and social media.
- Traditional database systems were designed to address smaller volumes of structured data, fewer updates or a predictable, consistent data structure.
- Big Data analysis includes different types of data

Handling bigdata- Parallel computing

- Imagine a 1gb text file, all the status updates on Facebook in a day
  - Now suppose that a simple counting of the number of rows takes 10 minutes.
    - Select count(*) from fb_status
  - What do you do if you have 6 months data, a file of size 200GB, if you still want to find the results in 10 minutes?
  - Parallel computing?
    - Put multiple CPUs in a machine (100?)
    - Write a code that will calculate 200 parallel counts and finally sums up
    - But you need a super computer

Handling bigdata - Is there a better way?

- Till 1985, There is no way to connect multiple computers. All systems were Centralized Systems.
  - So multi-core system or super computers were the only options for big data problems
- After 1985, We have powerful microprocessors and High Speed Computer Networks (LANs , WANs), which lead to distributed systems

- Now that we have a distributed system that ensures a collection of independent computers appears to its users as a single coherent system, can we use some cheap computers and process our bigdata quickly?

MapReduce Programming Model

- Processing data using special map() and reduce() functions
- The map() function is called on every item in the input and emits a series of intermediate key/value pairs(Local calculation)
- All values associated with a given key are grouped together
- The reduce() function is called on every unique key, and its value list, and emits a value that is added to the output(final organization)

Not just MapReduce

- Earlier count=count+1 was sufficient but now, we need to

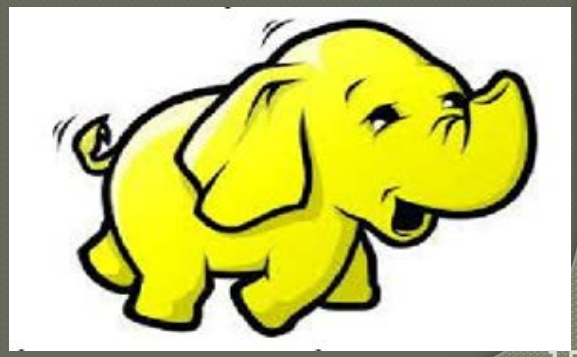
1. Setup a cluster of machines, then divide the whole data set into blocks and store them in local machines
2. Assign a master node that takes charge of all meta data, work scheduling and distribution, and job orchestration
3. Assign worker slots to execute map or reduce functions
4. Load Balance (What if one machine is very slow in the cluster?)
5. Fault Tolerance (What if the intermediate data is partially read, but the machine fails before all reduce(collation) operations can complete?)
6. Finally write the map reduce code that solves our problem

- Ok. Analysis on bigdata can give us awesome insights.
- But, datasets are huge, complex and difficult to process.
- I found a solution, distributed computing or MapReduce
- But looks like this data storage & parallel processing is complicated
- What is the solution?

Hadoop

- Hadoop is a bunch of tools, it has many components. HDFS and MapReduce are two core components of Hadoop
  - HDFS: Hadoop Distributed File System
    - makes our job easy to store the data on commodity hardwar
    - Built to expect hardware failures
    - Intended for large files & batch inserts
  - MapReduce
    - For parallel processing

- So Hadoop is a software platform that lets one easily write and run applications that process bigdata



Why Hadoop is useful

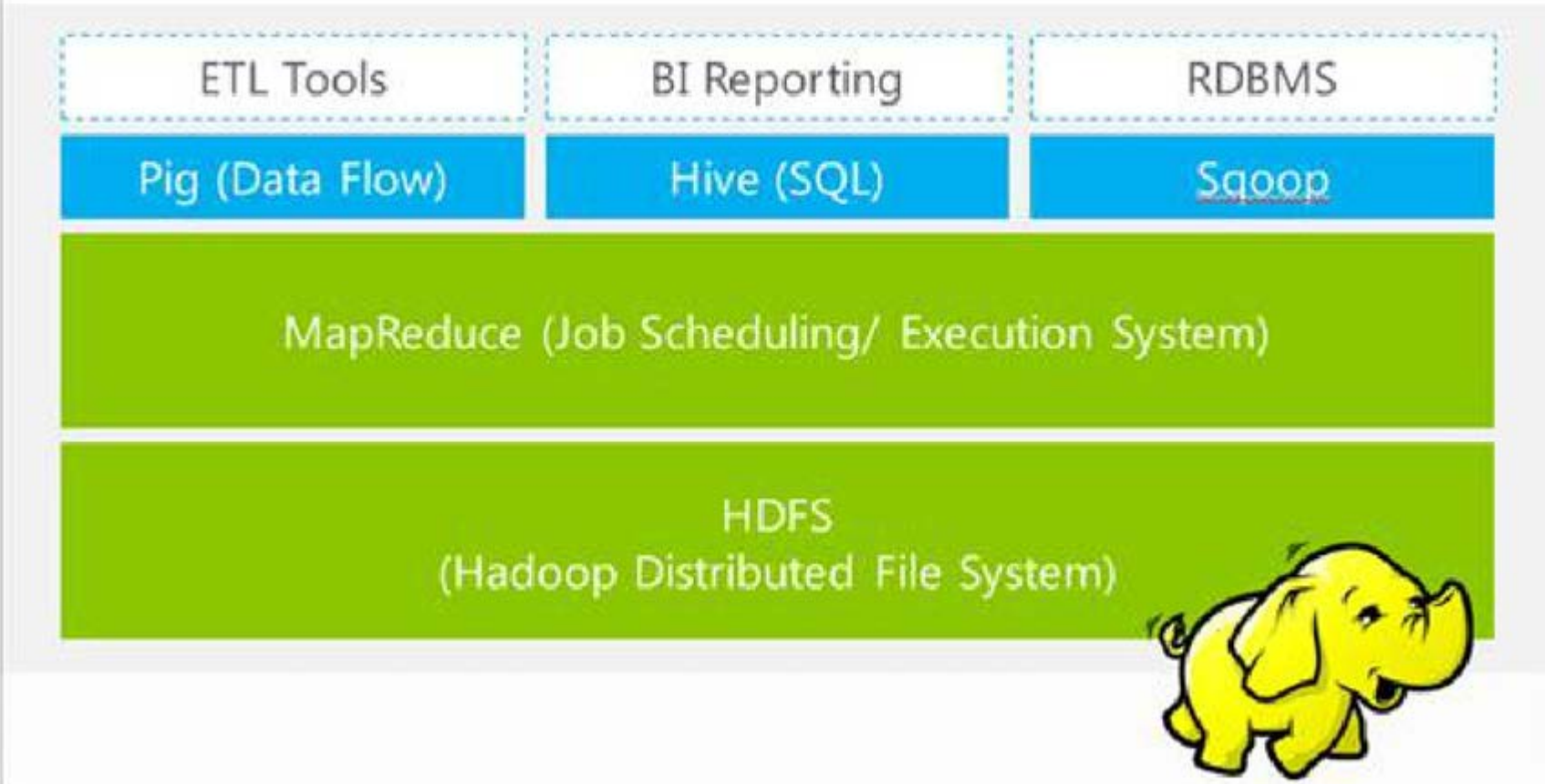
- **Scalable:** It can reliably store and process petabytes.
- **Economical:** It distributes the data and processing across clusters of commonly available computers (in thousands).
- **Efficient:** By distributing the data, it can process it in parallel on the nodes where the data is located.
- **Reliable:** It automatically maintains multiple copies of data and automatically redeploys computing tasks based on failures.
- And Hadoop is **free**

So what is Hadoop?

- Hadoop is not Bigdata
- Hadoop is not a database

- Hadoop is a platform/framework
  - Which allows the user to quickly write and test distributed systems
  - Which is efficient in automatically distributing the data and work across machines

Hadoop ecosystem



Big Data ecosystem

Big Data Analytics

- Examining large amount of data
- Appropriate information
- Identification of hidden patterns, unknown correlations
- Competitive advantage
- Better business decisions: strategic and operational
- Effective marketing, customer satisfaction, increased revenue

Types of tools used in Big-Data

- Where processing is **hosted**?
  - Distributed Servers / Cloud (e.g. Amazon EC2)
- Where data is **stored**?
  - Distributed Storage (e.g. Amazon S3)
- What is the **programming model**?
  - Distributed Processing (e.g. MapReduce)
- How data is **stored & indexed**?
  - High-performance schema-free databases (e.g. MongoDB)
- What operations are performed on data?
  - Analytic / Semantic Processing

Application Of Big Data analytics

Smarter Healthcare

Multi-channel sales

Homeland Security

Telecom

Traffic Control

Trading Analytics

Manufacturing

Search Quality

Risks of Big Data

- Will be so overwhelmed
  - Need the right people and solve the right problems
- Costs escalate too fast
  - Isn't necessary to capture 100%
- Many sources of big data is privacy
  - self-regulation
  - Legal regulation

Benefits of Big Data

- Our newest research finds that organizations are using big data to target customer-centric outcomes, tap into internal data and build a better information ecosystem.
- Big Data is already an important part of the $64 billion database and data analytics market
- It offers commercial opportunities of a comparable scale to enterprise software in the late 1980s
- And the Internet boom of the 1990s, and the social media explosion of today.

www.edutechlearners.com