A

PROJECT REPORT

ON

“BIG DATA”

As a Partial Requirement for the Degree of

BACHELOR OF COMPUTER APPLICATION

(B. C. A.)

Submitted to

C.B. PATEL COMPUTER COLLEGE &

J.N.M. PATEL SCIENCE COLLEGE,

BHARTHANA, VESU, SURAT

Affiliated to

VEER NARMAD SOUTH GUJARAT UNIVERSITY,

SURAT.

ACADEMIC YEAR: 2021- 2022

Guided by:

DR. AMI DESAI

Submitted by:

VIVEK J. VAGHASIYA

Agenda

- introduction of Big data
  - What is big data
  - History of big data
  - Advantages of big data
  - Disadvantages of big data
  - Big data Features
  - Handling big data-parallel computing
  - Why to learn big data

- Basic of Big data
  - Application of big data
  - Characteristic of big data
  - Big data web frame
  - Big data eco system
  - Big data analytics
  - Types of tools using big data
  - Why people use big data
  - Big data references

1. Introduction of Big data

1. What Big data

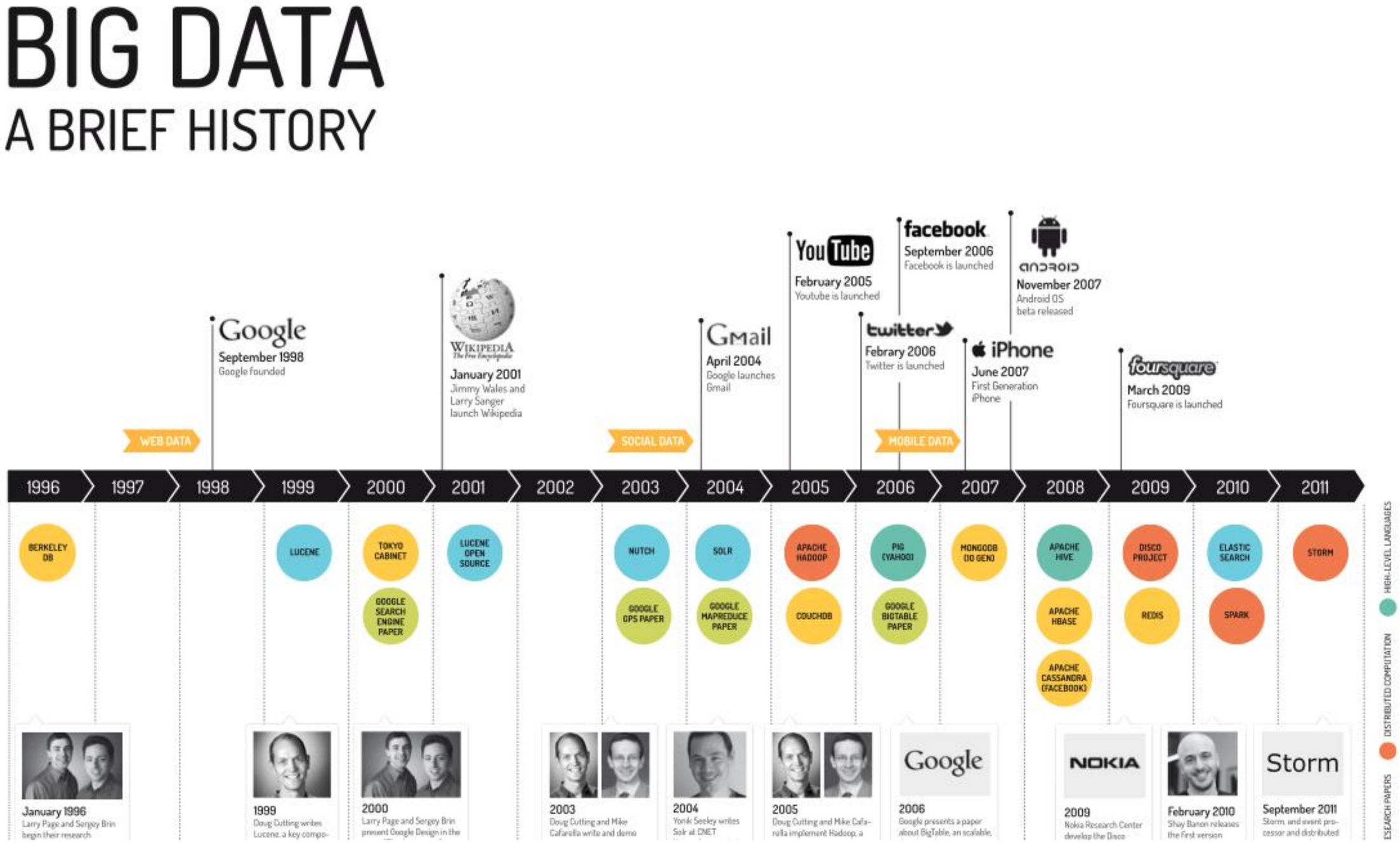
- Collection of data sets so large and **complex** that it becomes **difficult to process** using on-hand database management tools or traditional data processing applications.

- "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.

- Big Data generates value from the storage and processing of very large quantities of digital information that cannot be analyzed with traditional computing techniques.

History of Big data



- The tenns of 'Big Data' has been in use since the early 1990's.

- Although it is not exactly know who first used the term, most people credit John **R.mashey** for making the term popular.

- In 2005 Roger **Mougalas** from **O'Reily** Media coined the term Bid Data, only a year after they created the term Web 2.0.

- It refers to large set of data that is almost impossible to manage and process using traditional business intelligence tools.

Advantages 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.

Disadvantages 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

Big data features

- **Data processing**
- **Predictive Application**
- **Analytics**
- **Reporting**
- **Security**
- **Technologies Support**

Handling Big data- 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 row takes 10 minutes.

- Select count(*) from fb_status.

- What do you do if you have 6 months data, a file of size 200GB,lf you still want to find the results in 10 minutes?

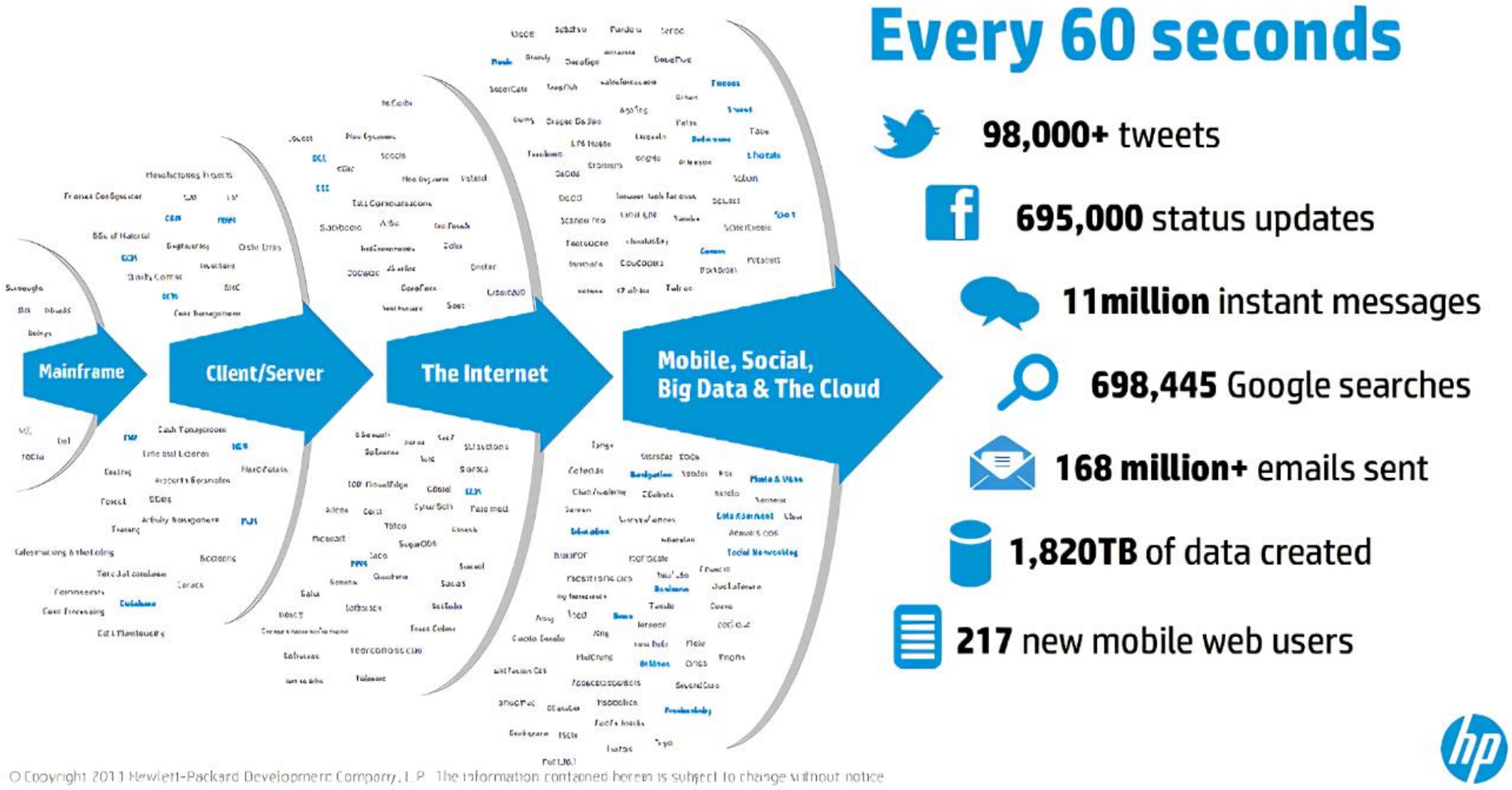
- Parallel computing?
- Put multiple CPUs in a machine (100?)
- Write a code that will calculate 200 parallel counts and finally sums up.
- But you need a super computer.

Why to learn Big data

- Data driven decisions provide a competitive Advantages.
- Big Data provides a spring for AI Artificial Intelligence (AI) is one of the most desired areas of expertise in business today.
- What most people not realize, however, is that Big Data provides a foundation for organizations that want to start AI projects.

- Big Data skills are in high Demand.

2. Basic of Big data



Application of Big data

- Smart healthcare
- Homeland security
- Traffic control
- Manufacturing
- Multi-channel sales
- Telecom
- Trading analytics
- Search quality

Characteristic of Big data

1st Character of Big Data Volume

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

- Today, Facebook in gests 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 billions of new, constantly-updated data feeds containing environmental,

location, and other information, including video.

2nd Character of Big Data Velocity

- Click streams and add 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

- Online gaming system support millions of concurrent user, each

producing multiple input 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 web Frame

- **Hadoop**
- **Map Reduce**
- **Spark**
- **Flink**
- **Storm**

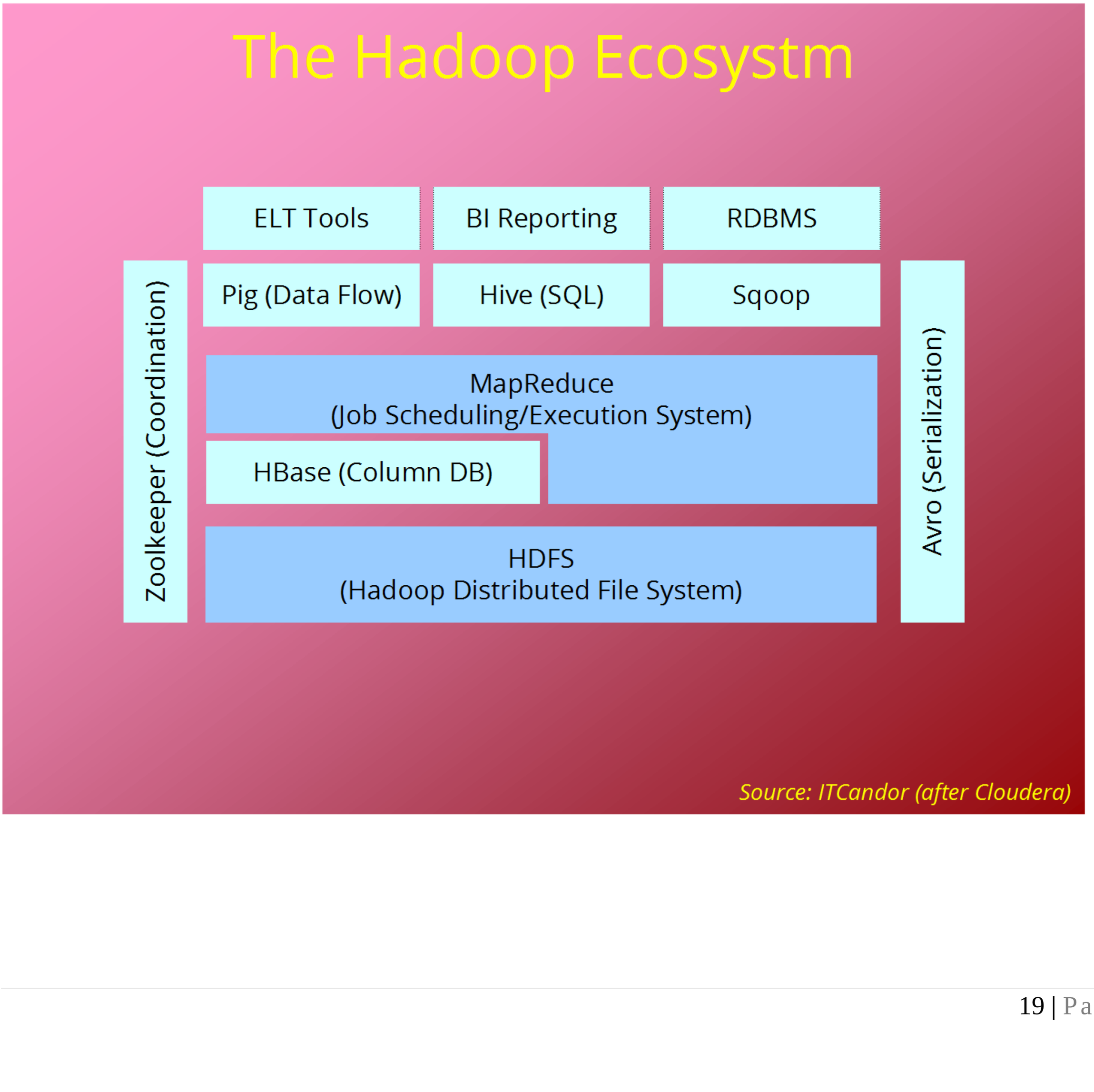
Hadoop

- Hadoop is a bunch of tools, it has Many components. HDFS and Map Reduce are two core components of Hadoop
  - HDFS: Hadoop Distributed File system.

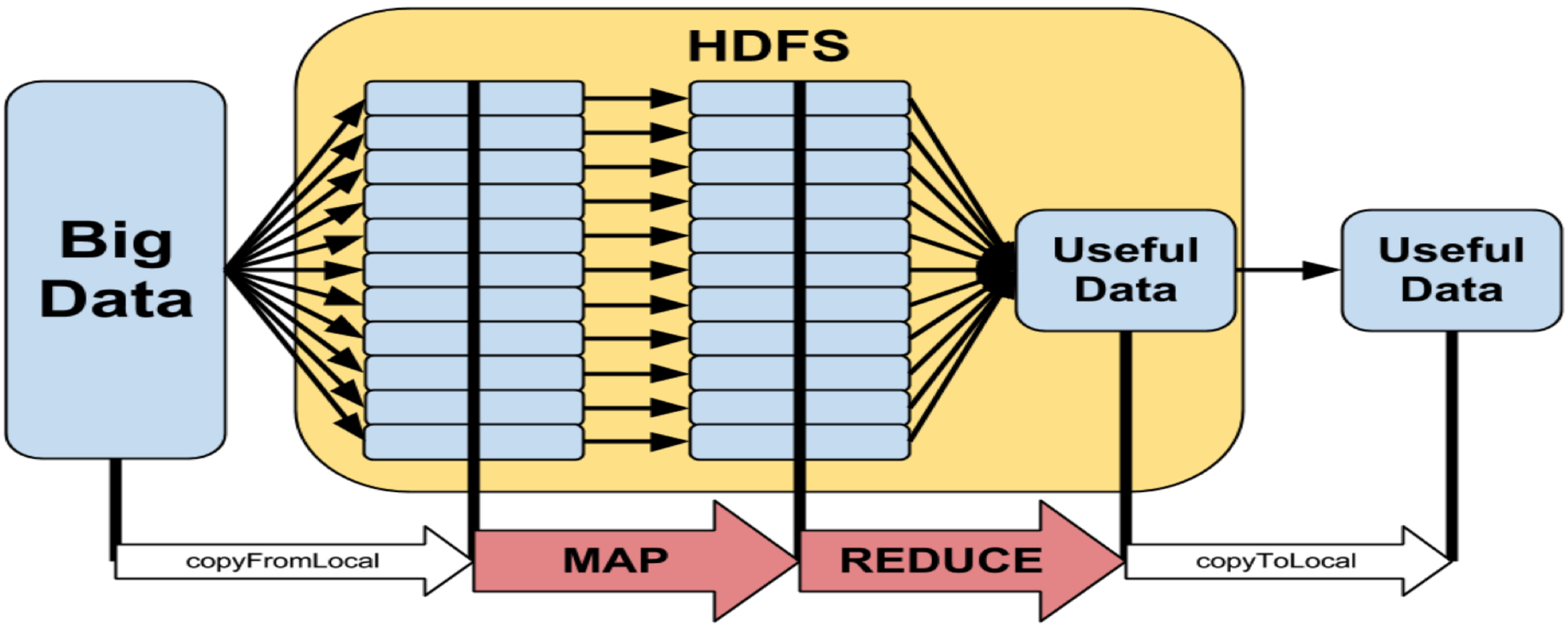
  - makes our job easy to store the Data on commodity hardware.
  - Built to expect hardware Failures.
  - Intended for large files & batch Inserts.
  - Map Reduce
  - For parallel processing

- So Hadoop is a software platform That lets one applications that Process big data.

Hadoop ecosystem



Map Reduce

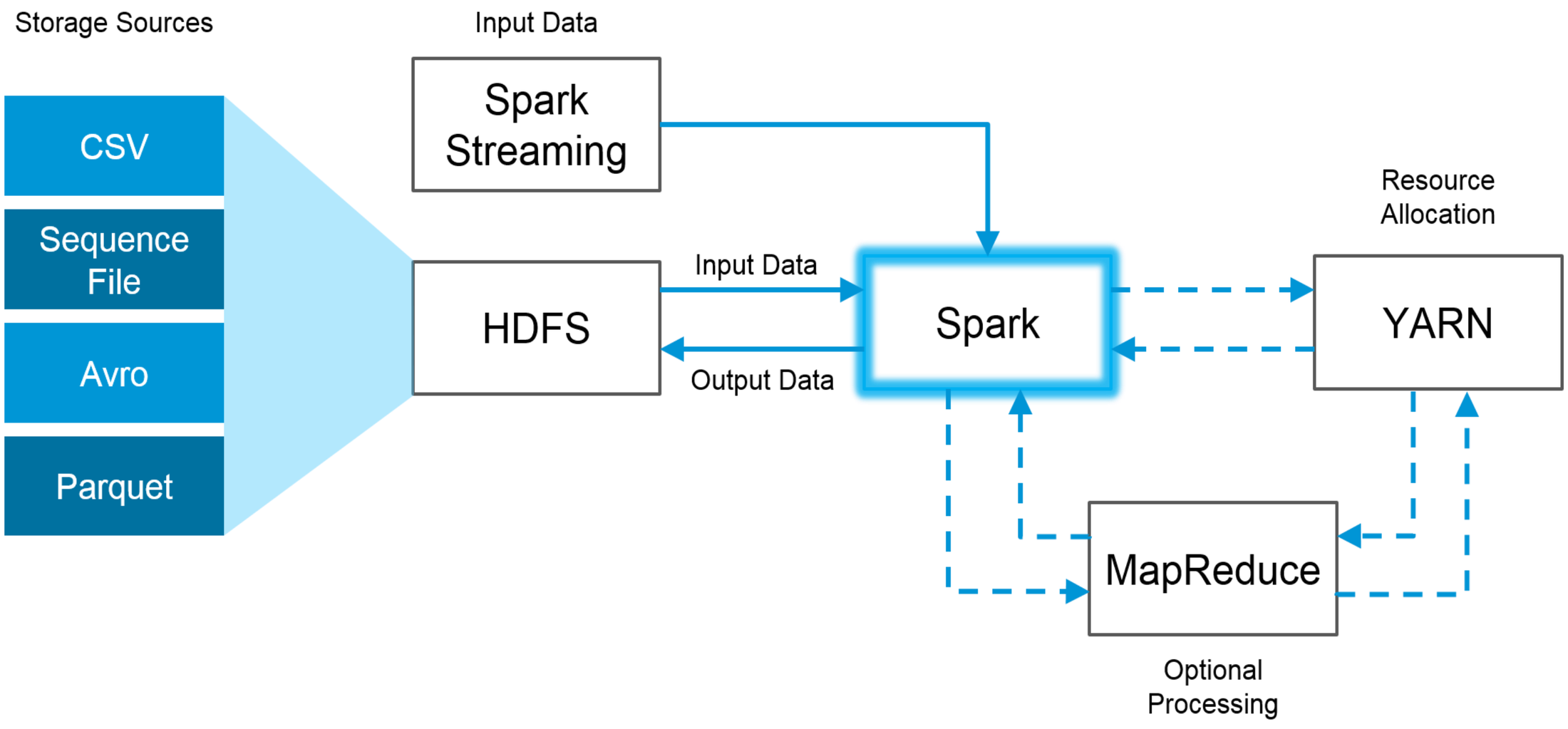


- Is this Big Data search engine getting outdated?

- The map() function is called on every item in the input and emits a series of intermediate key/value pairs(Local calculation).

- 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)

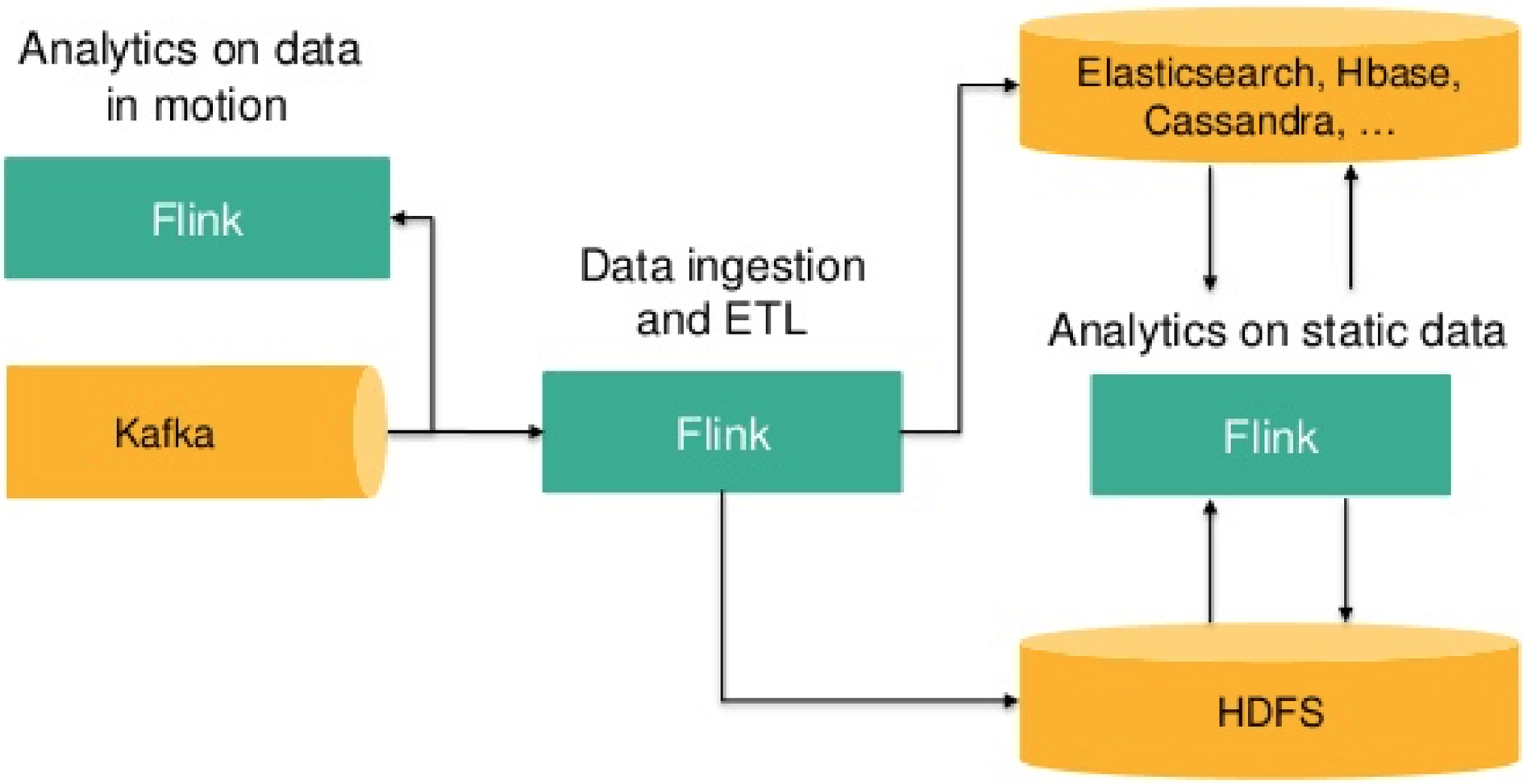
Spark



- Is it still that powerful tool it used to be?

- Fastest Batch processor or the most voluminous stream pocessor?

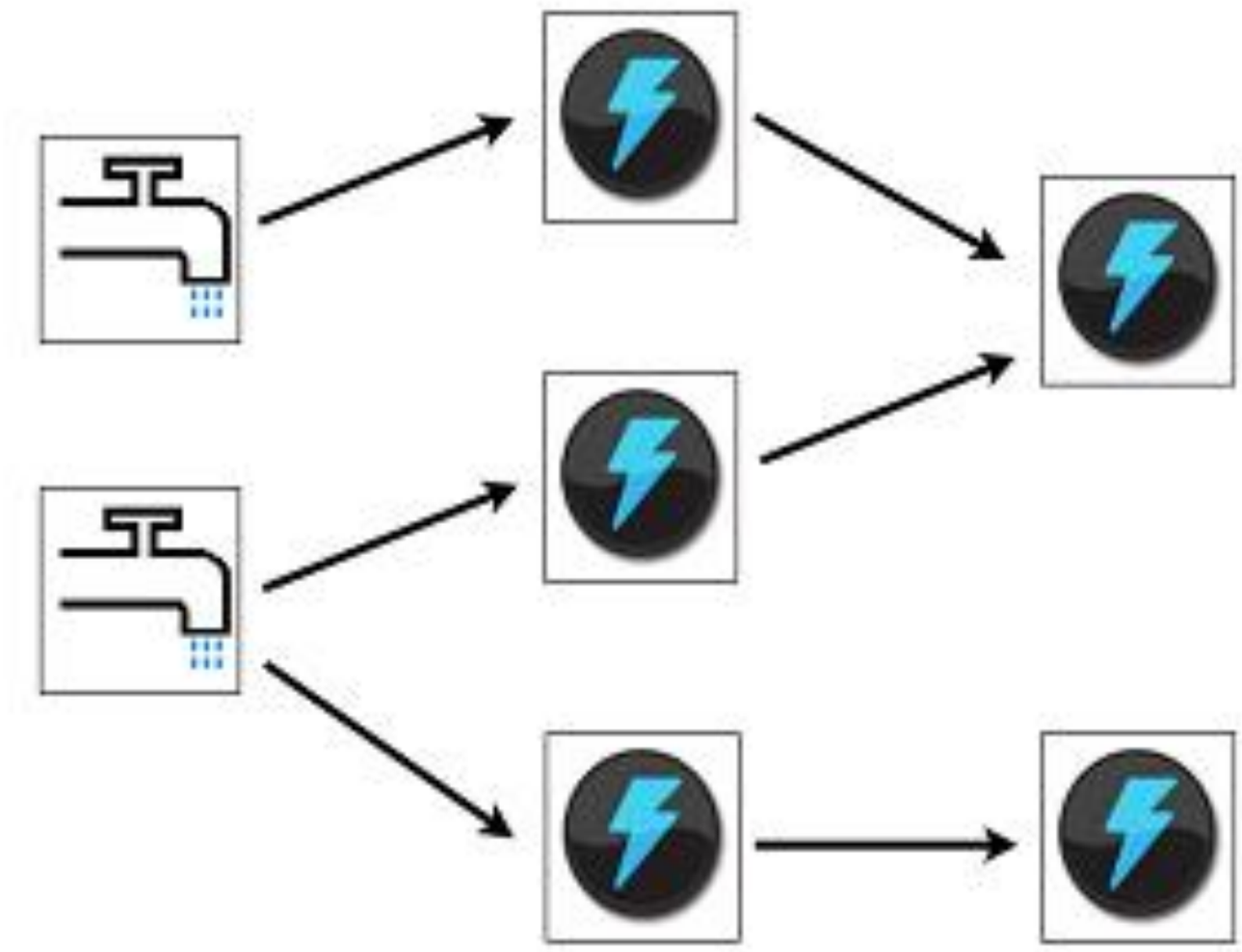
Flink



- **Apache Flink** is a streaming dataflow engine, aiming to provide facilities for distributed computation over streams of data.

- Flink is effectively both a batch and real-time processing framework, but one which clearly puts streaming first.

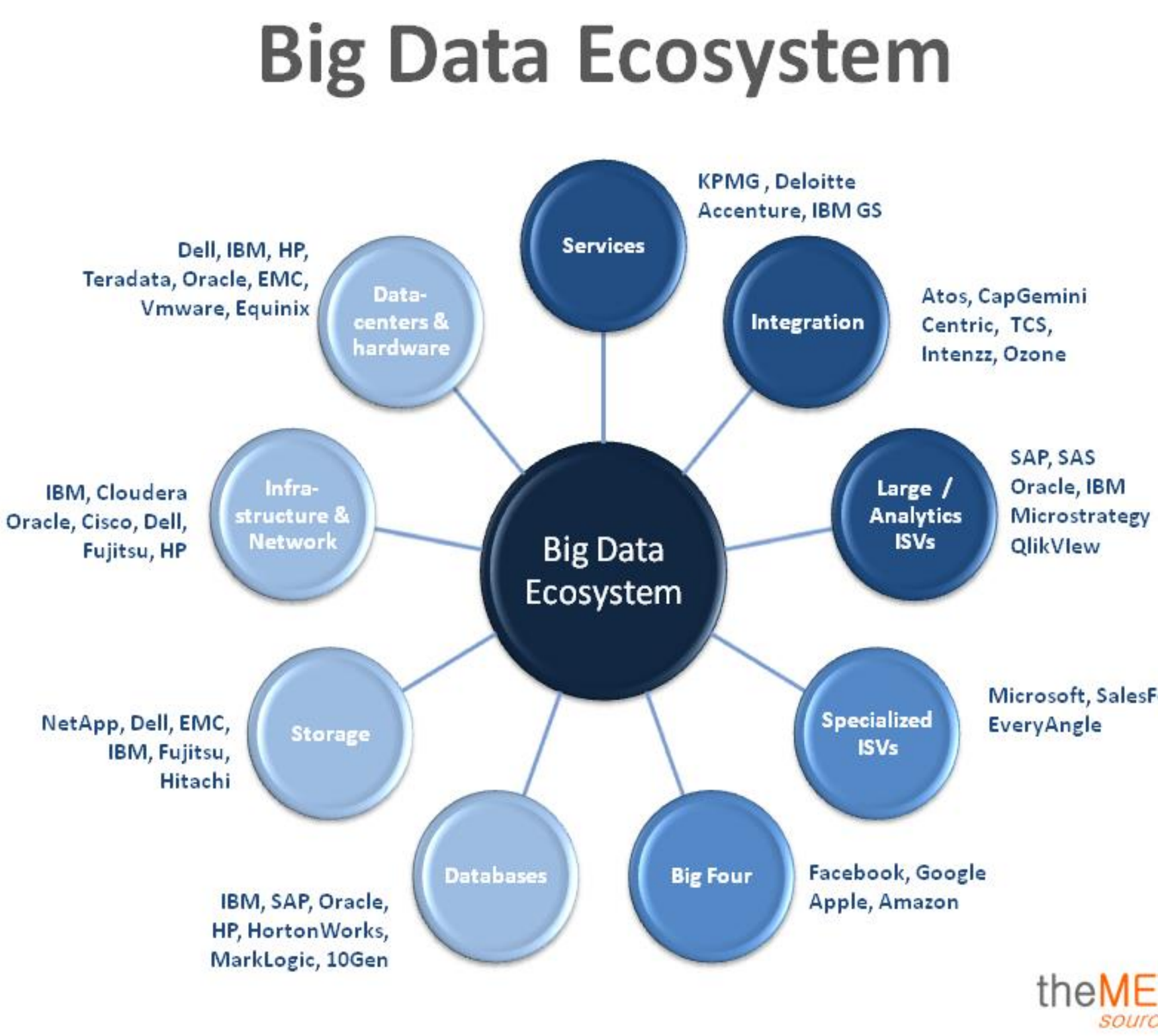
Storm



- **Apache Storm** is a distributed real-time computer system, whose application are designed as directed acyclic graphs.

- Storm is designed for easily processing unbounded streams, and can be used with any programming language.

Big data eco system



Big data Analytics

- The Big Data analytics is indeed a revolution in the field of Information Technology.

- The use of Data analytics by the companies is enhancing every year.

- The primary focus of the companies is on customers, Hence the field is flourishing in Business to Consumer (B2C) applications.

- We divide the analytics into different types as per the nature of the environment.

- We have three divisions of Big Data analytics: Prescriptive Analytics, Predictive Analytics, and Descriptive Analytics.

Types of tools use in big data

- Distributed Processing (e.g. MapReduce)

- How data is **stored** & **indexed?**

- High-performance schema-free databases (e.g. MongoDB)

- What operations performed on data?

- Analytic / Semantic processing.

- Hadoop - helps in storing and analyzing data;

- Talend - used for data integration and management;

Why people use Big data

- Big data analytics efficiently helps operations to become more effective.
- This helps in improving the profits of the company
- Bid data analytics tools like Hadoop helps in reducing the cost of storage.
- This further increases the efficiency of the business.
- The act of gathering and storing large amounts of information for eventual analysis is age old.

Conclusion

- The availability of Big Data, low-cost commodity hardware, and new information management and analytic software have produced a unique moment in the history of data analysis.

- The convergence of these trends means that we have the capabilities required to analyse astonishing data sets quickly and cost-effectively for the first time in history.

- These capabilities are neither theoretical nor trivial.

- They represent a genuine leap forward and a clear opportunity to realize enormous gains in terms of efficiency, productivity, revenue, and profitability.

- The Age of Big Data is here, and these are truly revolutionary times if both business and technology professionals continue to work together and deliver on the promise

= > Reference <=

- http://wikipedia.org/
- https://hadoop.apache.org/
- Google analytic:- one minute on internet

Thank You