Big Data processing framework – Hadoop in Data Science projects

By Manish Kalkar, Binay Jena, Rajasekhar Reddy Karna, Juan Guevara

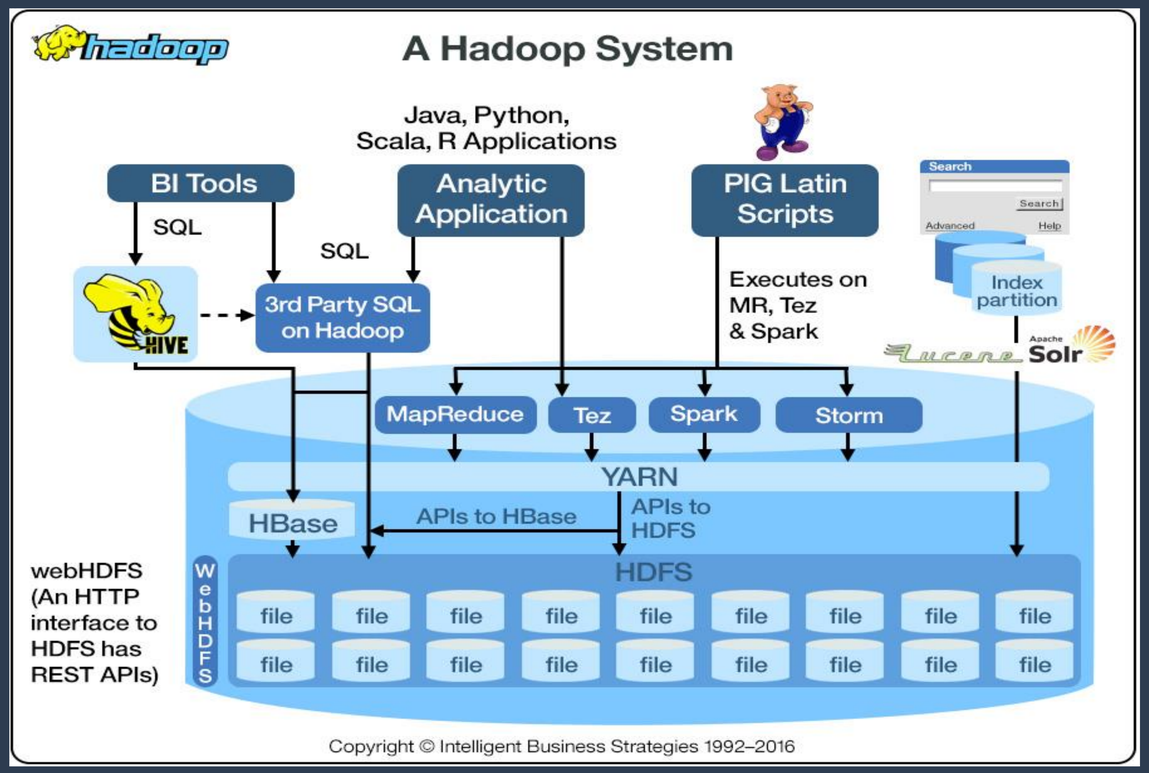
INTRODUCTION

Moore's Law[1] holds good five decades since its postulation - moving on from microprocessors we are now in the era of nanotechnology, smart devices, sensors, trackers, AI/ IoT devices. In recent decades, increasingly large amounts of data are generated from a variety of sources. The size of generated data per day on the Internet has already exceeded two exabytes. Within one minute, 72 hours of videos are up-loaded to YouTube, more than 100.000 Tweets are shared on Twitter and more than 200.000 pictures are posted on Facebook. Data driven decision making has become a routine affair for firms, business entities, governments, critical missions. The realization has set in that traditional relational databases and business intelligence solutions are inadequate to cater to this data driven decision-making needs. Questions for research (and implementation framework) are:

1. How do we make data storage infinitely scalable, while ensuring seamless retention, archival practices without affecting query retrieval and latency aspects?
2. How to ensure platform's compute abilities scale while ensuring optimal data processing performance ?
3. How does resource mapping, node allocation and execution tracking work in case of failovers in the framework?

In 2002, Apache Nutch project intended to create a search engine to index 1 billion web pages, which confirmed the real challenges at big data scale with scaling, implementation and maintenance costs. The Google File System[3] (2003) laid out the architecture details to deal with storage aspect of the problem, and MapReduce[2] (2004) paper did address the processing or computation aspect. Hadoop Framework - open sourced Java based application works in an environment that provides distributed storage and computation across clusters of computers, running on cheap/commodity grade hardware. Two major layers in Hadoop are (1) Processing/Computation layer: MapReduce is a parallel programming model for coding distributed applications for efficient processing of multi-terabyte datasets on large clusters of commodity hardware in a reliable and fault-tolerant manner. (2) Storage layer: provides a distributed file system designed to run on commodity hardware. In addition, Hadoop framework includes two modules, Hadoop Common - Java libraries and utilities and Hadoop YARN - framework for job scheduling and cluster resource management. Hadoop architecture becomes an automatic choice for data storage and processing platform due to its inherent fault / failover tolerance, scalability and interoperability / compatibility with traditional legacy systems.

Going beyond its original goal of searching millions of web pages and returning relevant results, Hadoop is being looked at as their next big data platform. Hadoop has become founding technology for Big data processing, Analytics, and Data Science! We dive deeper into the most widely used big data framework Hadoop, designed to cope with Big Data challenges.



WHY IS THIS TOPIC DATA SCIENCE?

Data Science in order to derive sense, actionable insights and patterns in data, requires interdisciplinary expertise in areas of computer programming, mathematics, statistics and industry domain. In execution, typical sequence of data science tasks involves storage, retrieval, cleansing, wrangling, serving – all of these being performed on large datasets on distributed computing to achieve efficiency.

Hadoop leverages distributed computing concepts through its storage (HDFS), scheduler and resource manager (YARN), and analyzer components (MapReduce), to perform all of the data handling tasks required in data science efforts. Additional modules like Pig, Hive lend further analysis muscle. Data science in essence is dependent on technology tasks performed on Hadoop ecosystem – this being the very tenet, 'bread and butter' operations of data science lends it relevance as must-know/must-have tool.

PROJECT DELIVERABLE

Enlightening data science enthusiasts with one of the must-haves in their artillery - Hadoop, to help in data science project execution. For a data scientist its necessary to know technology prowess and the way Hadoop components operate in terms of how's and whys of the tasks.

End Deliverable:

White paper with following topics, detailing the 'nuts and bolts' of Hadoop.

1. Introduction
2. Big Data Characteristics and Hadoop – Overview
3. Hadoop History and Timeline - Key Milestones
4. Key Features and Advantages of Hadoop
5. Hadoop Architecture - MapReduce, HDFS, YARN
6. Hadoop - Case Study
7. Conclusion
8. References

HADOOP ADVANTAGES

Hadoop Framework offer following benefits[4]:

- Flexibility: Hadoop stores data without requiring any preprocessing. Store data—even unstructured data such as text, images, and video—now; decide what to do with it later.
- Fault Tolerance: Hadoop automatically stores multiple copies of all data, and if one node fails during data processing, jobs are redirected to other nodes and distributed computing continues.
- Low Cost: The open-source framework is free, and data is stored on commodity hardware.
- Scalability: Hadoop system can grow, simply by adding more nodes.

CONCLUSION

We have often heard the platitude "A bad workman blames his tools". Hadoop empowers data science bypassing this cliché.

Parallel processing, distributed computing, optimal resource management, vertical and horizontal scalability, open sourced architecture, fault tolerance, custom analytics modules – all of these help reduce the technology friction aspects and lets data scientist focus on the business logic, core issue diagnosis, problem formulation and its resolution methodologies, thus letting the processing, workload management part of the execution be handled in the platform. It would be more than apt to mention Hadoop as the most prized weapon in the data scientist's artillery.

CASE STUDY

Flipkart[5], to stave off competitors Amazon, Paytm in Indian eCommerce space, needed cutting edge technologies at breakneck pace. Flipkart runs an inhouse data center with the scale of 5 PB RAM, 120 PB disk storage on Azure platform, where scaling is 'just' another non-functional requirement to handle promo events. Flipkart Data Platform (FDP) runs the show for metrics reporting, business insights, advanced analytics and personalization. FDP runs on 800-1000 node Hadoop cluster, storing 40+ PB data, ingesting up to 50+ TB daily, and processing 1.5 PB daily.

FDP Components powered by Hadoop Framework includes:

- FDP Ingestion System: Kafka Message queues, Dropwizard, HDFS, Quartz, Azkaban & Hive.
- Batch Compute Processing System: Hive or Vertica tables.
- Realtime Processing System: Apache Storm, Flink stack, Kafka.
- Data Lake: HDFS, Message Queue on Kafka topics.
- Knowledge Graphs: Metadata, and Data Lineage is powered through Apache Spark GraphX library.

LITERATURE

1. Gordon E Moore. 1965. Moore's Law. Retrieved from https://en.wikipedia.org/wiki/Moore%27s_law
2. Jeffrey Dean, Sanjay Ghemawat. 2004. MapReduce: Simplified Data Processing on Large Clusters. https://static.googleusercontent.com/media/research.google.com/en//archive/mapreduce-osdi04.pdf
3. Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung. 2003. The Google File System. https://static.googleusercontent.com/media/research.google.com/en//archive/gfs-sosp2003.pdf
4. Masters' In Data Science Org. Why use Hadoop? https://www.mastersindatascience.org/data-scientist-skills/hadoop/
5. Rishabh Dua. 2018. Flipkart Data Platform — India's largest eCommerce Big Data Platform - https://tech.flipkart.com/overview-of-flipkart-data-platform-20c6d3e9a196

ACKNOWLEDGEMENTS

Our sincere thanks to:

- Project Group 1 Team members for excellent teamwork
- DSC 500 Class for their encouragement
- Dr. Parajulee for providing constructive feedback
- Entire Hadoop research community for their inspiration