**The State and Future**

**of**

**Big Data Processing Platforms**

**By**

**Narava Pranay Manikanta**

**MT50255**

**DATA 603: Platform for Big Data Processing**

**Professor: James Klucar**

**May 02,2023**

**Abstract**

Big Data is most heard term in Business and Organizations. In this digital era, data became a crucial part for development of business and organizations. and big data processing platforms have become an essential tool for processing the data. This research paper explores the history and challenges of these platforms, as well as one upcoming big data platform. As Big data became a critical aspect of today’s businesses as Traditional data processing methods are no longer effective as data sets continuously increasing. traditional methods of managing data may struggle to manage the Volume, Variety, Velocity, Value and Veracity of data generated by modern systems.

This can lead to issues with scalability, data variety, processing speed, cost, and complexity, which can impact business operations and decision-making. To manage these types of issues several types of frameworks are developed in that most used platforms are Hadoop, spark, Dask these are the very well know big data platform throughout the industries these platforms are used for the processing of huge amount of data. Even though managing Big Data presents several difficulties, including the requirement for scalable storage and processing infrastructure, real-time data processing, and advanced data analysis. This paper aims to provide an overview of the evolution of Big Data processing platforms, discuss current challenges, and highlight an up-and-coming framework of Apache Flink.

**DATA:**

In today’s world data is everything. Analyzing a simple set of data give a beautiful and amazing insights to business. Data is not just limited to business, it is used in various fields like scientific research studies, Government policies and Health care. Data Became a Crucial part in every one life It helping organizations make informed decisions to improve products and services, identify trends, allocate resources, track student progress, and many more. As the data being generated the data base is continuously increasing, due to this new and innovative ways of using the data increased this leading to further advancement of various fields. As the data base increases, dealing with the data become more complex. Using traditional methods to process the data became exceedingly difficult because of the data velocity, volume, and its variety. This has leads to development of the Big Data Technology to process the data more efficiently.

**Big Data:**

Big Data is a term used to describe large and complex data collections that can neither be managed nor evaluated using traditional methods for processing data. In addition to the sheer volume of data, it also refers to the velocity at which the data was created as well as the variety, value, and veracity of the data.

**Volume**:

The huge amount of data produced by the digital technology. These data mostly in petabytes, and even in exabyte range for example social media and E-commers platforms generates terabytes of data every day in various forms like images, texts, videos, audios. The amount of data implies to the volume of the data. Using the tradition storage method to store this amount of data is difficult and at same time it was hard to retrieve the data. For these big data technologies has been introduced like Amazon s3, HDFS (Hadoop distribution file system)

**Velocity**:

Velocity is the speed of the data generated and processed with the increasing in data most of the sectors are depending on the real time stream processing. In today the data is generating in unprecedented pace for that we need a reliable and rapid processing and analysis technologies, traditional technology is unable to work with this rapid generation of data. For example, e commerce websites generate tons of date every minute these data need to be processed in real time and provide useful recommendations to the user.

**Variety**:

Variety of data means diverse data set. Which consists of structured, semi structured, and unstructured data for example a scientific team collect several types of data from various sources and this data is different from each other. like images, different equipment readings and other. For this advanced technology in big data need, to manage variety of data like Apache Cassandra, MongoDB

**Veracity**:

Veracity is the reliability and accuracy of the data. It is especially important when dealing with the data to provide insights and data driven decisions. For example, social media generated vast amounts of data both structured and unstructured. However social media also generates inaccurate and bias data like fake news, trolls, bots posts and many more. So, to ensure that the data is reliable and accurate the data need to be cleaned and transformed accordingly this process can be time taking and hard to process using traditional methods. For these big data technologies needed along with variety of API

**Value**:

Value of the Data refers to benefits and insight provided by the data example e-commerce website generate vast amounts of the data. with this data they can be able to analyze the patterns and trends of a customer behavior which helps e-commerce to recommend the products according to likes of the users. So, extracting useful insights and values from the data is important.

These are the 5 Vs in the Big Data. These characteristics can affect the data processing in many ways. For computing this large data set in big data we need powerful computing technologies such as Hadoop, Spark, Dask and other big data frameworks. In the processing of big data was categorized into five key steps they are data source and formats, data storage, data processing, data analysis and Data governance and security.

**Data sources and formats:**

Data Can be in various forms and formats like images, videos, text, or speech. This data can be either structured data, semi structured data or unstructured data.in the data is generated through various medium, in today’s world every digital technology generating ton of data very day from transactions to internet, social media, and many more.

**Data Storage:**

once the data was collected from various sources with diverse types of data formats. Then this data needs to be stored in either cloud database server or on-premises server. Since big data is so huge it often required distributed storge across the multiple servers. And most popular storage technologies are Amazon S3, HDFS (Hadoop Distribution File System) and Apache Cassandra. These are most popular and finite technologies using in the Big Data industries.

**Data Processing:**

To Analysis the Data we need to process the data first by transforming the Unstructured Data into the Structured Format. The traditional data processing techniques will not work on the Big Data due to its Volume, Velocity, Value, Variety, Veracity. To manage this big data the popular big data platforms are used like Hadoop, Apache Spark, Dask, Map reduce. These are the big data Frameworks that can manage large set of data and able to process and transform the data. And in this only some technologies can be able to manage both batch processing and real-time stream processing.

**Data Analysis:**

Once the data is transformed from raw unstructured data to structured data then it can be used for analyzing the data and visualization which used to help data driven decision making. Most popular data visualization tools are Power BI, Tableau.

**Data Governance and security:**

Once data is processed and analyzed we need to take appropriate Security measures to Secure the data. To Avoid the Cyder attacks.

**Evolution of Big Data:**

Hadoop was the first Big Data Framework introduced in 2005. It is an open-source framework used for distributed storage medium and big data processing. This Hadoop was based on the map reduce programming model to process the data in the distributed computing environment. But Hadoop lacks in speed and scalability this prompted the researcher and technologist to work on further development of big data frameworks that are more efficient. Later NoSQL database was introduced which provided easy access to store and retrieve the unstructured data.

And later, years cloud computing too developed this technology helped many companies to reduce their infrastructure and maintenance cost by providing data center that are remote to store and access the data. And one of the major changes happened when real-time stream process was introduced it made most of the companies’ jobs at ease. Real-time stream process helps for fraud detection and monitoring as the data received and processed in real-time. And later, year combination of both batch process and real time stream process has been introduced which helps organizations to process the data more dependable, flexible, and efficiently.

And, in addition to these technologies GPU accelerated Technologies also introduced which are Edge Computing technologies like Jetson Nano these technologies helping industries to transform rapidly. moreover, these evolution in technologies are driven for the need of more dependable, efficient, scalable, and real-time processing of the data. These technologies helping organizations to process and analyze the data as it generates and in more innovative ways.

**Future of Big Data:**

Big Data has evolving rapidly since 2005 from the first launch of Hadoop. The processing of big data becoming increasingly fast but at the same time Data also becoming more complex and continue to grow. In this evolution of big data one of the trends moving towards the real-time streaming data processing. This type of process gaining more popularity due to it processes the data as it generates. This type of data processing technology helping companies to provide real-time analysis better and accurate insights.

In this type of processing Apache Flink gaining a more popularity compared to other frameworks. Initially Flink was starts development in 2009 and first stable release was in 2016. Flink is best known for its Real-time Stream processing. this is first framework truly developed for real-time stream processing. Flink also does support Batch processing, Graph processing, Iterative processing too. As the big data Evolves Flink became a next generation engine for the real-time stream processing. As we know spark is faster than Hadoop but experts believing Flink is much faster than Spark due to Flink’s low latency and high throughput applications. and Flink providing Robust fault-tolerance Using the State & Checkpointing feature which means upon failure of any task it can resume where it left of. And Flinks recallability helps to add more features while running the tasks although maintaining exactly one semantic. Which means every record is processed only once’s and will not be repeated.

Even though Apache Spark can perform Real-time Stream processing but by heart it was best known for batch processing. Initially spark was designed for Batch Processing later it was converted to support real-time stream processing too. sparks stream computational model was based on Micro batching which means it divide the data into discrete chunks of data and then it processes the data, this is known as micro batches. Spark is mimicking stream process by treating the data into number of small batches so that spark is considered as near to real-time stream processing engine. Where Flink uses the concept of windowing and checkpoint in data stream where it splits the data into buckets with finite size.

And, Spark does not have an efficient Memory manager. Instead, it has custom memory manager where user must do memory management task manually unlike Spark, Flink has its own Automatic memory manager, and it provides manual configuration too. Spark frequently encounter out of memory error due to incompetent memory manager. Where Flink experience exceedingly rare out of memory errors. Even though whenever memory about to run out it swaps the memory with disk and Flink supports Controlled cyclic dependency graph as it is execution engine this makes Flink to represent machine learning in efficient manner this help to improves the overall performance in processing the data.

Along with it Flink has a dedicated rich set of libraries and API which makes Flink Easy to model complex real-world problems. These helps to development of Graph processing, Machine learning, and String handling. And Flink provides two Relational API which are Table API and sequel API. In terms of performance Flink is best known for outperforming the spark in real-time stream processing. Due to this Flink become the most trending topic in the big data community due to its real-time stream processing. in near future we can see even much more advanced frameworks which helps to process the data more efficiently and helps to provide valuable information in real-time as the data generates.

Now coming to architecture Flink does not have its own storage system even though it is distributed data flow engine, it uses external storage facilities/technology like Amazon S3, HDFS (Hadoop distribution File System), Kafka, Cassandra. And Flink is integrated with cluster resource management like Kubernetes, YARN and other commonly used clusters and it also can be used as standalone cluster too. And it can run stateful streaming at any scale. Flink Consist of both job manager and task manager. Scheduling the jobs, management of jobs and execution plans will be done using job manager by allocating required resources. Where Task manager helps to execute user defined functions.

**Conclusion:**

In Conclusion the Big data is the fastest evolving technology in the digital world from the launch of Hadoop in 2005. Within two decades of time the big data rapidly grow into advanced technology to process the data in real time. This technology evolves more in upcoming years. I believe this Flink will guide the upcoming technologies to advance further. These technologies help to process and analyze the data more efficiently. As the 5V’s of data velocity, volume, Variety, veracity, value of data generated by modern system can be able to process using the Big Data Technologies. And Flink providing the major support to process the complex data in real time and analyze the data as it generated. The Apache Flink is faster than the Spark. As Flink having a many Advantages like Scalability, bounded and unbounded processing, Stateful stream processing, API support, optimization, Flexible file formats, flexible deployments, Fault tolerance, low latency, and high throughput. This has made the Flink most popular framework as a real-time stream engine available today. Flink can manage big datasets. And many big companies are using Flink like Uber, Alibaba and many more. As well as Flink also have few limitations like limited API Support, Basic Machine learning Support. Apache Flink is a powerful low latency streaming engine than can handle all type of workload. And one more big data trend is moving towards the cloud processing. The big data technology is becoming more advance to process the data at ease.

# References

*Apache Flink - Architecture*. (n.d.). Retrieved from Tutorialspoint: https://www.tutorialspoint.com/apache\_flink/apache\_flink\_architecture.htm

*Apache Flink® — Stateful Computations over Data Streams #*. (n.d.). Retrieved from Apache Flink: https://flink.apache.org/

Chintapalli, S. a. (2016). Benchmarking Streaming Computation Engines: Storm, Flink and Spark Streaming. *2016 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, 1789-1792.

Diego García-Gil1\*, S. R.-G. (2017). A comparison on scalability for batch big data processing on Apache Spark and Apache Flink.

GĘBIŚ, W. (2022, 11 29). *What is Apache Flink? Architecture, Use Cases, and Benefits*. Retrieved from Nexocode: https://nexocode.com/blog/posts/what-is-apache-flink/

Hanif, M. a. (2018). An Efficient Topology Refining Scheme for Apache Flink. *2018 International Conference on Information and Communication Technology Convergence (ICTC)*, 766-770.

Katsifodimos, A., & Schelter, S. (2016). Apache Flink: Stream Analytics at Scale. *2016 IEEE International Conference on Cloud Engineering Workshop (IC2EW)*, 193-193.

Paris Carbone, A. K. (2015). Apache Flink™: Stream and Batch Processing in a Single Engine. *IEEE Data Eng. Bull.*, 28-38.