

Reference Data Architecture For Extreme Data Processing In Real Time

Arun Uniyal

Himalayan Garhwal University

Dept. of Computer Science & Applications

Accenture Solutions Pvt. Ltd.

Manager

Gurugram, Harayana, India

Abstract—

- i) Data which is known to be key asset for any enterprise has pivotal role in driving businesses decisions.
- ii) The technology advancement has open new avenues which has transformed the way we generate, consume or use data to derive value from it.
- iii) The speed at which data is generated is growing rapidly, which is quite complex, leading to advancement in the field of big data. “Big data” is a term used for massive data sets which have complex structure and cannot be handled by standard software or platforms.
- iv) The category of data which has huge potential to become part of this big data journey, huge computational needs and usage over the coming years will be due to introduction of IoT (internet of things) based devices, digital economies, digital devices and connected devices.
- v) This category of data has possibilities due to Digital Transformations and will be empirical to look for possibilities where it can be effectively used with a faster turn-around to consumers for their analytical or reporting needs.
- vi) This paper will try to explore alternative frameworks which we can use for evolving data giving more computational power to existing Big Data Platforms highlighting why they are important given the current challenges with available platforms, tools or technologies.

Keywords — Analytics Platforms, Artificial Intelligence, Big Data, Blockchain, Cloud Analytical Platforms, Data Architecture, Data Engineering, Data Lake, Digital, Ensembling, Hyperparameters, In-Memory, Internet of Things (IoT), Machine or Deep Learning, Master Data Management, Statistical Modelling

Date of Submission: 02-08-2023

Date of Acceptance: 12-08-2023

I. INTRODUCTION

Data has become part of everybody’s journey today. The pace at which we generate, store and consume data across channels is humongous. With reducing cost of hardware, organizations are leveraging their real potential to store data which is used for business decisions across variety of generated content. Technological advancement has further helped to make best use of semi-structured and unstructured data sets, which has left some food of thought to enterprises to refocus on business from a different perspective. Many organizations have transformed their strategy to become data driven where they are using the derived value from it to take business decisions. The Data Architecture by virtue of this has taken shape with emerging Big Data based platforms with limitations to process massive data sets in real time. Though aggregated data exists in the platform but at times, there is a need to engineer the massive data set which is not possible in current ecosystem.

II. DATA PLATFORMS

With evolution of modern data platforms, which includes big data & cloud analytical platforms, businesses had to rethink how they want to process their data. On the one side, these platforms have managed to ease out storage and processing complexities to a bit by providing end to end solution for analytics problem focusing it precisely from an infrastructure or platform perspective.

Big Data is a term which has picked pace in the recent years. The platform which offers big data capabilities addresses the following problems around data:

1. **Volume:** As compared to traditional databases, big data platforms are capable of handling huge volume of data pipes. The volume of data may range from terabytes to exabytes depending on the use case. Systems

like click stream, social media feeds, IoT devices, digital feeds etc. generate huge datasets every set which may derive values to organizations.

2. **Velocity:** It defines the speed at which data is received from sourced and an action is taken on it. Data coming from some of the sources like IoT devices or applications required immediate evaluation. Examples may relate to recent instances where data from smart wearables helped to save lives of individuals when acted immediately.
3. **Variety:** Data types other than structured data like semi-structured and unstructured data like Video, Audio requires different types of analysis, tools so that they are reasonably processed. Some of commonalities they share with structured data are lineage, metadata management, auditable etc. As data comes from different source, it adds complexity how it is treated and tacked back to sources for any business need.
4. **Veracity:** It refers to trustworthiness of data. As big data processes data coming from variety of source and types, it is important that it maintains a quality, timeliness and accuracy. Until these parameters are met it is hard to control them.

With all these key aspects, the platform at times address how it will evolve over time to derive value from it. Unless it is delivering value, it makes no sense to implement such a system.

III. CLOUD DATA INFRASTRUCUTURE

Cloud is term used for collection of connected networks. It takes away all the complexity of setting up data centers locally and help to maintain infrastructure by exposing them as services. Hence the requirements for local hardware and software requirements to meet any business need are reduced by shifting load to network provider over internet. The three key services offered by cloud computing are:

1. Infrastructure as a Service (IaaS)
2. Platform as a Service (PaaS)
3. Software as a Service (SaaS)

The benefits which cloud has to offer are:

1. Elastic infrastructure
2. On-Demand service
3. Reduced Capital & Ongoing costs
4. Availability
5. Quality of service
6. Choice of tools

Cloud computing in context of big data, offer both storage and computing capabilities to address core capabilities. As the requirements around data processing or storage varies, the cloud infrastructure ensures that they are always met by shifting load from actual consumer to external cloud partner. One of core reason behind evolution of cloud infrastructure is considerable reduction in storage costs. In addition, adoption of pay-as-you-go model and use of commodity hardware has helped to evolve big data as a service.

IV. ANALYTICS PLATFORM REFERENCE ARCHITECTURE

With increasing focus on data over the last few years, organizations want to reap benefits of data accumulated over years. With evolution of big data and cloud data platforms, some standards are set forward with industry acceptance of the architecture which is not only proven but has set a landmark in terms of baseline. The widely accepted reference data architecture in the current modern age is:







No	Notation	Description
1		The notation of a data store or database which represents structured, unstructured, semi-structured or real-time (streaming) data. We often indicate a short text or description of the notation.
2		This indicates the flow between two notations in a single direction. The left arrow of the line indicates the flow of the connection occurs from left to right notation.
3		This notation represents Data Processing and Computation. We often indicate a short text or description of the notation.
4		This represents data storage in file formats.
5		This indicates the flow between two notations in both directions.
6		This represents collection of processes which together supports the core platform or infrastructure.

Fig 1: Reference architecture notations

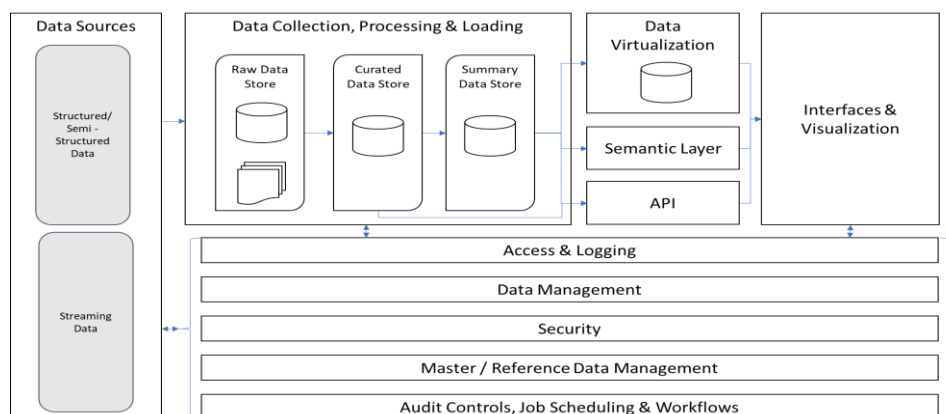


Fig 2: Big Data Reference Architecture

1. **Data Sources:** It refers to the ecosystem which is generating data and is received by big data platform from them. This includes all varieties of data.
2. **Data Collection, Processing & Loading:** This layer in the big data platforms transforms the data into business usable KPI's. This layer is also used to translate data to a format which analysts or data science experts can use to derive some value.
3. **Data Virtualization:** This refers to technology which helps to merge data coming from various sources in order to generate insights on the fly.
4. **Semantic Layer:** It is a conceptual layer which enables abstraction between underlying data storage and representable KPI's in reports, dashboards or visualizations.
5. **Interface & Visualizations:** These applications consume data generated by big data platform as downstream applications. At times, KPI's are presented or data is used for further analysis / translation.

V. CHALLENGES WITH CURRENT ANALYTICAL PLATFORMS

The current analytical platforms on-premise or on cloud has limitations which we need to overcome in order to tackle developments in this area. Some of the key challenges are listed below:

1. **Data Management:** Managing and governing data in analytical platforms is one of the biggest challenges given the nature of data sources. Heterogenous sources of data is pumped at near real time fashion into platforms leading to quality and trust issues. Lack of data management over the time leads to traditional data challenges starting from performance, maintenance, trust, quality, governance and above all usability.
2. **Effective use of platform:** While we talk about big data platforms, the core resources used are storage and processing power for huge data sets. If not managed properly, the platforms will mostly be used to dump data then deriving value from it. In case of cloud instances, it not only leads to overspending than an in-premise platform but wastage of resources from both storage & processing perspective leading to overspending, exposure to vulnerabilities, identification of key information from humongous data accumulated over time, audit trail issues, and at times un-intentional in-adherence to compliance or regulatory requirements.
3. **Limitations to address modern age data application needs:** With evolution of artificial intelligence based applications, machine or deep learning, traction of businesses to use larger data sets to derive value leveraging statistical models, the current infrastructure fails to address the ask by limiting the data sets or with limited processing capabilities which if scaled to huge training data sets or entire population of data may lead us nowhere. We need to rely on smaller sample size in order to derive some conclusions, ensemble models at times or invest heavily on tuning hyperparameters to achieve that smaller shift of confidence to next level.
4. **Processing Sensor based, Unstructured or Semi-Structured data sets:** The platforms available since ages can address the need to process structured data sets but with advancement in technology, majority of the data generated today comprise of unstructured or semi-structured data sets. On top of this, smart digital devices have opened avenues to process sensor-based data which will eventually increase with further adoption in the industry of IoT based applications like smart homes, smart grids, connected cars etc. The data assumed to be generated from these applications is anticipated to grow many folds which may need further food for thought in order to analyze it real time.
5. **Architectural limitations:** The current industry adopted reference data architecture has many capabilities with limitations to address ever changing need to analyze massive data sets in real time. The big data platforms available today on premise or on cloud can store and process data in real time but when it comes to analyze large data sets in real time, they fall behind. The architecture should evolve where it should not only process such huge data sets but produce insights in a timely fashion to derive insights and take decisions.

Industries like Pharma, Telecom, Finance have many applications where massive data should be processed real time to derive value. One of the latest examples is outburst of Coronavirus where large processing engines may be required to analyze both research and patient data to derive conclusions. With limited processing capabilities from available platforms, the scientists will have to rely on small samples to deduce research outcomes.

VI. PROPOSED REFERENCE ARCHITECTURE

The proposed reference architecture will address the processing needs of emerging technological data trends. As one of the biggest challenges in current platforms is to process massive data sets, the reference architecture will evaluate how we can drive value from large data sets in real time. The reference architecture will use existing framework to retrofit gaps.

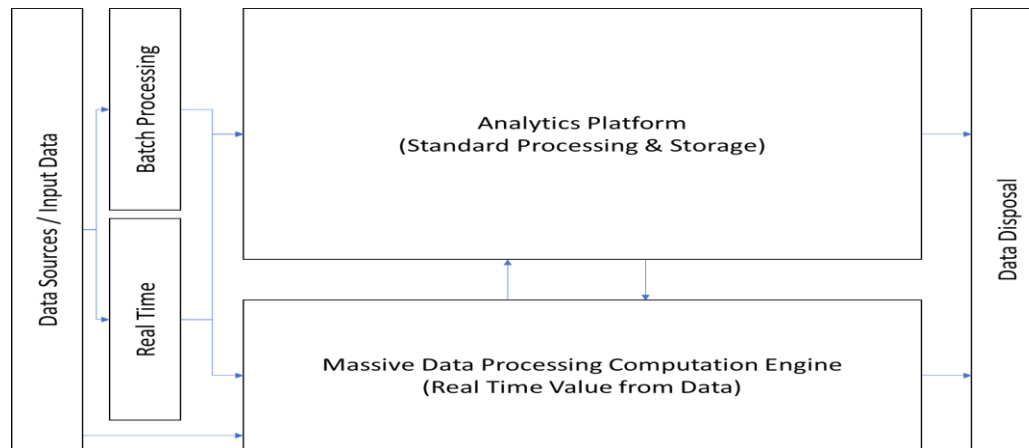


Fig 3: Proposed Reference Data Architecture

VII. USAGE OF PROPOSED REFERENCE ARCHITECTURE

The proposed architecture is based on the shortcoming from existing widely accepted big data reference architecture specifically for processing power. In past many attempts were made to enhance this capacity ranging from In-Memory engines to massive parallel processing but none of them has potential to analyze streaming data in large volumes that too in real time. Further to this, added complexities of data manipulations like transformations, aggregations, statistical engineering etc. tops up capabilities which are limited in the current landscape. With the proposed architecture, it will benefit in the following ways:

1. **Speed:** The additional framework exposed in parallel to analytics platform will offer potential to speed up processing power. The already available processing power of the platform can be used in parallel to reuse what is available with temporarily using the storage exposed by already setup analytical platform. This will ensure that when need be, the requirements to analyze massive data sets be processed via parallel route and take conclusive decisions on the fly without storing data. While the decisions may be spontaneous, it required engine which can process data at a speed which is quite faster to already available processing power. At times, likewise the use cases of Machine Learning, Pharma etc. may need high potentials than collective processing power offered by platforms.
2. **Ease of Use:** The proposed architecture considers the widely accepted big data reference architecture as a foundation, hence adoption with end users specially in the context of data consumption will be easy. The existing reference platform is extended beyond existing capabilities hence it will be quite easy to scale the use cases on the parallel processing engine.
3. **Change Enablement:** The organizational change which is required in adoption of newly proposed architecture is minimal. This change only requires minimal training, primarily for IT staff or likes of Data Scientists, Analysts etc. who may potentially use real time data for analysis or decision making.
4. **Potential Reduction in Storage Requirements:** As the proposed architecture focuses on real time data analytics, the temporary data treatments or storage may not be required in contrary to traditional mechanisms, where data is prepared, stored and then used for consumption.
5. **Use of existing Governance & Lake:** The reference architecture thus proposed uses existing governance framework with least modification leveraging data from sources or data lake formed so far by the organizations across use cases depending on how they want to leverage historical data with real time analytics.

VIII. FUTURE AND BEYOND

This paper tries to address the processing limitations of big data platforms in context of processing capabilities required for real time analytics across huge data sets. With technological advancements in the field of data architecture, lot of efforts are invested to look at how problems across data supply chain can be resolved to address business need. Ever evolving technologies like Artificial Intelligence, Machine Learning, IoT Platforms, Digital adoption will lead us to an era where data engineering and analytics will face challenges to derive real time value given the need of the hour. With businesses relying more on data, traction have moved towards real time data analytics than delaying the decisions to reap benefits from data right at the point where data is generated or bring better customer experience. Many organizations have failed in past as they were not able to timely use data for decision making. More and more businesses are now relying heavily on data with transition to data driven organizations from business-driven ones. It is predicted that data will fuel the growth if used properly.

The future though looks bleak with many global challenges and weak economies, but the traction and investments for transformations towards data centricity has increased many folds over the past few years. There are businesses who are benefitted from deriving data value in a timely fashion by managing data as an asset. With the current trends on how industry can leverage data should focus on some of the following challenges to rightly use data for business decisions:

1. **Data Management:** Right management of data will lead to timely actions which will benefit the businesses in many ways e.g. customer adoption, business growth etc.
2. **Data Quality:** There are still challenges while capturing data in the ecosystem where if not captured properly, the final outcomes are way apart from what is anticipated. If based on wrong data, the decisions will adversely affect the outcomes. Hence managing the right quality of data will not only ensure growth but outcomes which are guaranteed from the analytics over this data set.
3. **Master Data Management:** The reference architecture though assumes “Master Data Management” as support pillar, but to address the key actors of the businesses, it is essential to manage it properly.
4. **Real time data analytics:** As cited above, traction for real time analytics has increased over time. The current infrastructure available has certain limitations to address this problem, if data grows beyond a certain volume. Hence adoption to new architecture is required in order to derive value from data with volumetric in real time.

The options explored above will lay foundation to how we can address the challenges from an architecture perspective. Even though, this will require extensive work and research, it enables to discover key elements, challenges, techniques and methods, hence would allow us making better reference architecture.

IX. CONCLUSION

This paper used industry accepted reference big data platform to look at limitations from processing perspective for real time data analytics for large data sets. The challenges for real time data analytics was evaluated in the existing architecture, which was evolved to next level to propose an alternative processing engine in order to bring data analytics in real time. There is still some work needed to look for options where technology or systems should be overlaid on the proposed reference architecture to solve real time problems, but the architecture hence proposed will lay a foundation to use to evaluate alternatives which can be evolved over time to set it as a base. The proposed architecture is recommended based on real time industry challenges and work with industry leaders, which was acknowledged and need adoption with evolving technologies to solve problems.

REFERENCES

- [1] Carl Matthew Dukatz, San Jose, CA (US); Daniel Garrison, Washington, MI (US); Lascelles Forrester, Conyers, GA (US); Corey Hollenbeck, Arlington, MA (US), “United States Patent Publication, No: US 2018 / 0308000 A1”, Oct 2018
- [2] Mrs. Bharati M. Ramageri, DATA MINING TECHNIQUES AND APPLICATIONS, ISSN : 0976-5166
- [3] Danijela Efnusheva, Ana Cholakovska And Aristotel Tentov, “A SURVEY OF DIFFERENT APPROACHES FOR OVERCOMING THE PROCESSOR-MEMORY BOTTLENECK,” In International Journal Of Computer Science & Information Technology (IJCSIT) Vol 9, No 2, April 2017.
- [4] K. Siddardha, Ch. Suresh, “Big Data Analytics: Challenges, Tools And Limitations” International Journal Of Engineering And Technical Research (IJETR), ISSN: 2321-0869 (O) 2454-4698 (P), Volume-6, Issue-3, November 2016.
- [5] Go Muan Sang, Lai Xu, Paul De Vrieze, “A Reference Architecture For Big Data Systems” Conference Paper From Researchgate.Net At <https://www.researchgate.net/publication/272027871> .
- [6] Pekkapääkkönen And Danielpakkala, “Reference Architecture And Classification Of Technologies, Products And Services For Big Data Systems” Researchgate.Net At <https://www.researchgate.net/publication/272027871> .
- [7] Samiya Khan, Kashish Ara Shakil And Mansaf Alam, “Cloud Based Big Data Analytics: A Survey Of Current Research And Future Directions”, Researchgate.Net At <https://www.researchgate.net/publication/281144737> .