Volume 5, Issue 5, May 2015 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Overview on Real-Time Big Data Analytics: Emerging Architecture

[1]K. Devika Rani Dhivya*, [2]Vijaya Nilaa.V

[1]M.Sc., M.Phil., M.B.A., Asst.Professor, Dept of CA&SS, Sri Krishna of Arts and Science College, Tamil Nadu, India

[2]II BCA A, Dept of CA&SS, Sri Krishna of Arts and Science College, Tamil Nadu, India

Abstract: Few years ago, analysts working with big datasets made queries and got the results back. The hadoop and other tools made it possible to get the results from queries in minutes. Analysts now demand sub-second, near real-time query results. Fortunately, we have the tools to deliver them. This report examines tools and technologies that are driving real-time big data analytic (RTBDA)s. This may help for those looking to keep up with current technology trends. A perfect balance between factual models, analytics and philosophy, this paper gives you theoretical insights into the most pertinent big data advancements taking place today.

Keywords- BigData, RTBDA stack, Five Phases Of Real Time Big Data.

I. INTRODUCTION

Real-time big data isn't just a process for storing petabytes or exabytesof data in adata warehouse. The co-author of Big Data, Big Analytics is Michael Minelli. It's about the ability to make better decisions andtake meaningful actions at the right time. It's about detecting fraudwhile someone is swiping a credit card, or triggering an offer while ashopper is standing on a checkout line, or placing an ad on a websitewhile someone is reading a specific article. It's about combining andanalyzing data so you can take the right action, at the right time, andat the right place. For some, real-time big data analytics (RTBDA) is a ticket to improvedsales, higher profits and lower marketing costs. This paper contains about how fast,how real,how real time big data analytics, RTBDA stack, five phases of real time.

What Is Big Data?

Big data analytics is the process of examining large data sets containing a variety of datatypes – i.e.,big data – to uncover hidden patterns, unknown correlations, market trends, customer preferences and other useful business information. This can lead to more effective marketing,new revenue opportunities, better customer service, improved operational efficiency, competitive advantages over rival organization and other business benefits. The primary goal of big data analytics is to help companies make more informed business decisions by enabling data scientists, predictive modelers and other analytics professionals to analyze large volumes of transaction data, as well as other forms of data that may be untapped by conventional business intelligence(BI) programs. That could include Web server logs and Internet clickstream data,social media content and social network activity reports, text from customer emails and survey responses, mobile-phone call detail records and machine data captured by sensors connected to the internet of things. Some people associate big data with semi-structured and unstructured data of that sort, but consulting firms like Gartner Inc. and forrester Research Inc. also consider transactions and other structured data to be valid components of big data analytics application. Big data which admittedly means many things to many people is no longer confined to the realm of technology. Today it is a business priority, given its ability to profoundly affect commerce in the globally integrated economy. In addition to providing solutions to long-standing business challenges, big data

Inspires new ways to transform processes, organizations, entire industries and even society itself. Yet extensive media coverage makes it hard to distinguish hype from reality – what is really happening? 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.

II. PERFORMANCE OF RTBDA

A. How fast is RTBDA

The capability to store data quickly is normal. The specialty is the capability to do something meaningful with that data, quickly and costeffectively. Businesses and governments have been storing huge amounts of data for decades. It is an explosion of new techniques for analyzing those large data sets. We're also find a proliferation of new technologies designed to handle complex, non-traditional datagenerated by social media, mobile communications, customer service records, warranties, census reports, sensors, and web logs. Heterogeneity is the new normal, and modern data scientists are accustomed to hacking their way through tangled clumps of messy data culled from multiple sources. Software frameworks such as Hadoop and MapReduce, which support distributed processing applications across relatively inexpensive commodity hardware, now make it possible to mix and match data from many disparate sources.

Today's data sets aren't like the older data sets they're significantly more complex. Every day it seems that a new technique or application is introduced that pushes the edges of the speed-size envelope even further. Druid, for example, is a system for scanning tens of billions of records.

B. How real is RTBDA

The meaning of "real time" can vary dependingon the context in which it is used.In the same sense that there really is no such thing as truly unstructureddata, there's no such thing as real time. There's only near-realtime,when we're talking about real-time or near real-time systems, what wemean is architectures that allow you to respond to data as you receiveit without necessarily persisting it to a database first.Real-time denotes the ability to process data as it arrives,rather than storing the data and retrieving it at some point inthe future. Fromthe point of view of an online merchant, "the present" means the attentionspan of a potential customer. If the processing time of a transactionexceeds the customer's attention span, the merchant doesn't considerit real time.From the point of view of an options trader, however, real time meansmilliseconds. From the point of view of a guided missile, real time meansmicroseconds.For most data analysts, real time means "pretty fast" at the data layerand "very fast" at the decision layer.

C. *How big is RTBDA*

As suggested earlier, the "bigness" of big data depends on its locationin the stack. At the data layer, it is not unusual to see petabytes andeven exabytes of data. At the analytics layer, you can encountergigabytes and terabytes of refined data. By the time you reachthe integration layer, you're also handling megabytes. At the decision layer,the data sets have dwindled down to kilobytes, andwe're measuringdata less in terms of scale and more in terms of bandwidth.The takeaway is that the higher you go in the stack, the less data youneed to manage. At the top of the stack, size is considerably less relevantthan speed.

III. THE RTBDA STACK

This paper is about an overview of the essential elements of "real time" big dataanalytics, and does a good job of explaining different approaches and various techniques. It should help to clear some of the confusion that occurs around the concept of big data and clarify the content of big data analytics within the statistical analysis. One and only goal of this paper is sketching out a practical RTBDA roadmap that will serve a variety of stakeholders including users, vendors, investors, and corporate executives such as CIOs, CFOs and COOs who make or influence purchasing decisions around information technology. Focusing on the stakeholders is very important and also their needs because it reminds us that the RTBDA technology exists for a specific purpose: creating value from data. We should also remember that "value" and "real time" will suggest different meanings to different subsets of stakeholders. There is presently no one-size-fits-all model, which makes sense when you consider that the interrelationships among people, processes and technologies within the RTBDA universe are still evolving. David Smith writes a popular blog for Revolution Analytics on open source R, a programming language designed specifically for data analytics. He proposes a four-layer RTBDA technology stack. Although his stack is geared for predictive analytics, it serves as a good general model: At the foundation is the **data layer**. At this level you have structured data in an RDBMS, NoSQL, Hbase, or Impala; unstructured data in HadoopMapReduce; streaming data from the web, social media, sensors and operational systems; and limited capabilities for performing descriptive analytics. Tools such as Hive, HBase, Storm and Spark also sit at this layer. (MateiZaharia suggests dividing the data layer into two layers, one for storage and the other for query processing) The **analytics layer**sits above the data layer. The analytics layer includes a production environment for deploying real-time scoring and dynamic analytics; a development environment for building models; and a local data mart that is updated periodically from the data layer,situated near the analytics engine to improve performance.

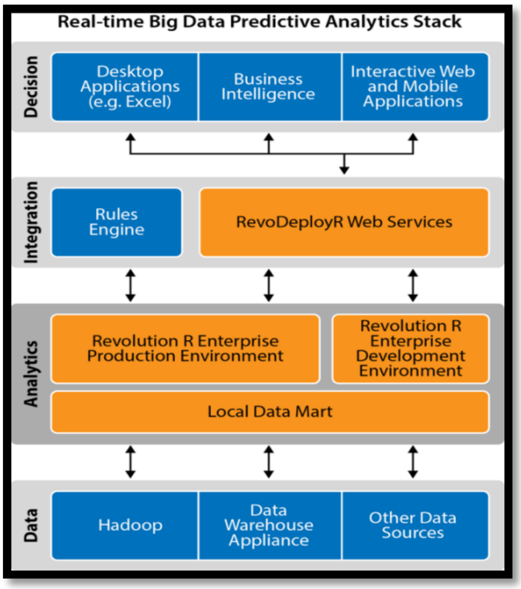
Real-time Big Data Predictive Analytics Stack
Decision
Desktop Applications (e.g. Excel)
Business Intelligence
Interactive Web and Mobile Applications
Integration
Rules Engine
RevoDeployR Web Services
Analytics
Revolution R Enterprise Production Environment
Revolution R Enterprise Development Environment
Local Data Mart
Data
Hadoop
Data Warehouse Appliance
Other Data Sources

Fig. 1 Real-time Big Data Predictive Analytics Stack.

On top of the analytics layer is the **integration layer**. It is the “glue” that holds the end-user applications and analytics engines together, and it usually includes a rules engine or CEP engine, and an API for dynamic analytics that “brokers” communication between app developers and data scientists. The topmost layer is the **decision layer**. This is where the rubber meets the road, and it can include end-user applications such as desktop, mobile, and interactive web apps, as well as business intelligence software. This is the layer that most people “see.” It’s the layer at which business analysts, c-suite executives, and customers interact with thereal-time big data analytics system. Again, it’s important to note that each layer is associated with different sets of users, and that different sets of users will define “real time” differently. Moreover, the four layers aren’t passive lumps of technologieseach layer enables a critical phase of real-time analytics deployment

IV. THE FIVE PHASES OF REAL TIME

Real-time big data analytics is an iterative process. This involves multiple tools and systems. There are five phases in RTDBA which called as Smith’s five phase process. Those phases are data distillation, model development, validation and deployment, real-time scoring, and model refresh. At each phase, the terms “real time” and “big data” are fluid in meaning. The definitions at each phase of the process are not carved into stone. Indeed, they are context dependent. Like the technology stack discussed earlier, Smith’s five-phase process model is devised as a framework for predictive analytics. But it also works as a general framework for real-time big data analytics.

1. Data distillation- Like unrefined oil, data in the data layer is crude and messy. It lacks the structure required for building models or performing analysis. The data distillation phase includes extracting features for unstructured text, combining disparate data sources, filtering for populations of interest, selecting relevant features and outcomes for modeling, and exporting sets of distilled data to a local data mart.

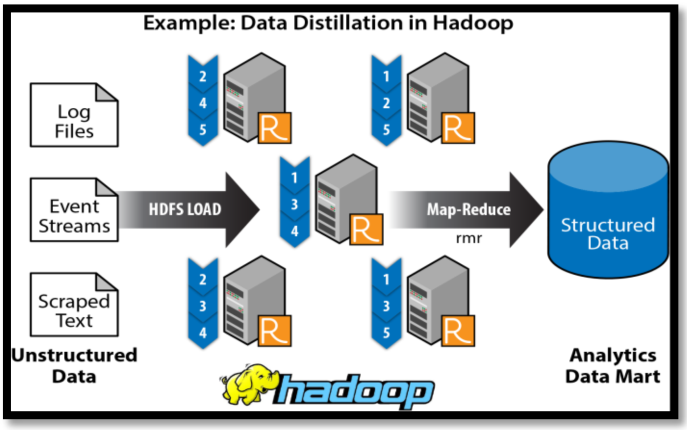

Fig. 2 Data Distillation in Hadoop

2. Model development- Processes in this phase include feature selection, sampling and aggregation; variable transformation; model estimation; model refinement; and model benchmarking. The goal at this phase is creating a predictive model that is powerful, robust, comprehensible and implementable. The key requirements for data scientists at this phase are speed, flexibility, productivity, and reproducibility. These requirements are critical in the context of big data: a data scientist will typically construct, refine and compare dozens of models in the search for a powerful and robust real-time algorithm.

3. Validation and deployment- The goal at this phase is testingthe model to make sure that it works in the real world. The validationprocess involves re-extracting fresh data, running it againstthe model, and comparing results with outcomes run on data that’sbeen withheld as a validation set. If the model works, it can bedeployed into a production environment.

4. Real-time scoring- In real-time systems, scoring is triggered byactions at the decision layer (by consumers at a website or by anoperational system through an API), and the actual communicationsare brokered by the integration layer. In the scoring phase,some real-time systems will use the same hardware that’s used inthe data layer, but they will not use the same data. At this phaseof the process, the deployed scoring rules are “divorced” from thedata in the data layer or data mart. Note also that at this phase,the limitations of Hadoop become apparent. Hadoop today is notparticularly well-suited for real-time scoring, although it can beused for “near real-time” applications such as populating largetables or pre-computing scores. Newer technologies such as Cloudera’sImpala are designed to improve Hadoop’s real-time capabilities.

5. Model refresh- Data is always changing, so there needs to be away to refresh the data and refresh the model built on the originaldata. The existing scripts or programs used to run the data andbuild the models can be re-used to refresh the models. Simpleexploratory data analysis is also recommended, along with periodic(weekly, daily, or hourly) model refreshes. The refresh process,as well as validation and deployment, can be automated usingweb-based services such as RevoDeployR, a part of the RevolutionR Enterprise solution.

VI. CONCLUSION

Today, most of our technology infrastructure is not designed for real time. Our legacy systems are geared for batch processing. We store data in a central location and when we want a piece of information, we have to find it, retrieve it and process it. That's the way most systems work. But that isn't the way the human mind works. Human memory is more like flash memory. We have lots of specific knowledge that's already mapped - that's why we can react and respond much more quickly than most of our machines. Our intelligence is distributed, not highly centralized, so more of it resides at the edge. That means we can find it and retrieve it quicker. Real time is a step toward building machines that respond to problems the way people do. As information technology systems become less monolithic and more distributed, real-time big data analytics will become less exotic and more commonplace. The various technologies of data science will be industrialized, costs will fall and eventually real-time analytics will become a commodity.

REFERENCE

[1] Mike is coauthor of The Executive's Guide to Enterprise Social MediaStrategy (Wiley, 2011) and Partnering with the CIO: The Future of ITSales Seen Through the Eyes of Key Decision Makers (Wiley, 2007).http://oreilly.com/catalog/errata.csp?isbn=9781449364212 for release details.

[2] LaValle, Steve, Michael Hopkins, Eric Lesser, Rebecca Shockley and Nina Kruschwitz. http://935.ibm.com/services/us/gbs/thoughtleadership/ibv-embedding-analytics.html

[3] Big data analytics –RTDBA stack safaribooksonline ,libraryhttps://www.safaribooksonline.comn/library/view/real- time- big data/ch05.html.

[4] definition o0f big data-analytics search business analytics definition http://searchbusinessanalytics./techtarget.com/definition/big-data-analytics.