Volume 1, Issue 2 (2013) 36-39 ISSN 2347 - 3258
International Journal of Advance Research and Innovation

Big Data and Distributed Data Mining: An Example of Future Networks

Prashant Kumar [b,*], Khushboo Pandey[a]

[a] Department of Computer Science & Engineering, DSITM, Ghaziabad, Uttar Pradesh, India
[b] Department of Computer Science &Engineering, Shri Venkateshwara University, Gajraula, U. P, India

Article Info

Article history:
Received 1 August 2012
Received in revised form
22 August 2013
Accepted 28 August 2013
Available online 20 September 2013

Keywords
Big Data,
Networks,
Business intelligent,
Apache Hadoop,
Apache MapReduce

Abstract

This paper describes the perspective on the analytics of big data generated by sensors and devices on the edge of networks. The paper includes a discussion of the importance of data at the edge of networks where some of "biggest" big data is generated. Also quick overview of emerging technologies, including distributed frameworks such as the Apache Hadoop framework and Apache* Map Reduce.

1. Introduction

The explosion of big data is testing the variety [1] [5], and velocity of this flood of complex, capabilities the explosion of big data is testing the

Fig: 1. Big Data

*** Corresponding Author,**
E-mail address: Prashant.iftn@gmail.com

variety [1] [5], and velocity of this flood of complex, capabilities of even the most advanced analytics tools. IT is challenged by the sheer volume, structured, semi structured, and unstructured data which also offers organizations exciting opportunities to gain richer, deeper, and more accurate insights into their business.

1.1. What is Big Data?

Big data is a buzzword, catch-phrase, used to describe a massive volume of both structured and unstructured data that is so large that it's difficult to process using traditional database and software techniques[6] [7]

Fig: 1.1. Big Data

Big data is typically described by the first three characteristics. The term big data is believed to have originated with Web search companies who had to query very large distributed aggregations of loosely-structured[23] data.

Big data analytics requires capturing and processing data where it resides. This paper explores the value of data at the edge of networks, where some of "biggest" big data is generated. As the use of sensors and devices as well as intelligent systems [4] [5] [6] continues to expand, the potential to gain insight from the flood of data from these sources becomes a new and compelling opportunity. Businesses that can harness the power of big data at the edge and unlock its value to the organization will outperform their competitors with greater capabilities to innovate creatively and solve complex problems whose solutions have been out of reach in the past. Below-sometimes referred to as the three Vs. However, organizations [6] [7] [12] need a fourth—value—to make big data work.

- Volume. Huge data sets that are orders of magnitude larger than data managed in traditional storage and analytical solutions. Think petabytes instead of terabytes.
- Variety. Heterogeneous, complex, and variable data[23][31], which are generated in formats as different as e-mail, social media, video, images, blogs, and sensor data—as well as "shadow data" such as access journals and Web search histories.
- Velocity. Data is generated as a constant stream with real- time queries for meaningful information to be served up on demand rather than batched.
- Value. Meaningful insights that deliver predictive analytics for future trends and patterns from deep, complex analysis based on machine learning, statistical modeling, and graph algorithms. These analytics go beyond the results of traditional business intelligence querying and reporting.

1.2. An Example of Big Data?

(The Apache Hadoop Framework and MapReduce) New technologies are emerging to make big data analytics possible and cost-effective [31]. The Apache Hadoop* framework is evolving as the best new approach. The Hadoop framework redefines the way data is managed and analyzed by leveraging the power of a distributed grid of computing resources.

The Hadoop open-source framework [5] [6] [7] [21] uses a simple programming model to enable distributed processing of large data sets on clusters of computers. The complete technology stack includes common utilities, a distributed file system, analytics and data storage platforms, and an application layer that manages distributed processing, parallel computation, workflow, and configuration management. In addition to offering high availability, the Hadoop framework is more cost- effective for handling large, complex, or unstructured data sets than conventional approaches, and it offers massive scalability and speed.

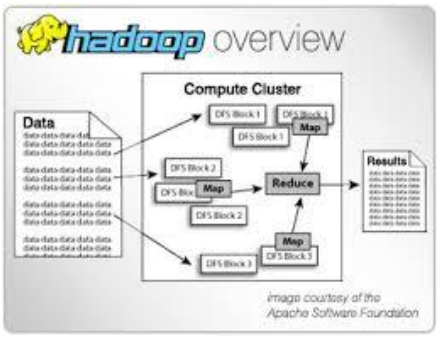

Fig: 1.2. The Apache Hadoop Framework

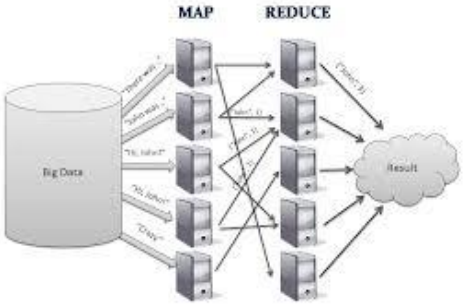

Fig: 1.3. MapReduce

2. Big Data at the Edge

Much of the current discussion about big data analytics today focuses on managing and analyzing unstructured data from business and social sources such as e-mail, videos, tweets, Face book posts, reviews, and Web behavior. While this type of big data analytics promises to provide significant value to organizations, data generated at the edge of the network from sensors and other devices represents another huge, untapped resource with the potential to deliver insights that can transform the operations and strategic initiatives of public and private sector organizations.

Data from intelligent systems and sensors is some of the largest volume, fastest streaming, and/or most complex big data. The data sources are distributed across the network and data is collected

by an enormous variety of equipment, such as utility meters, traffic and security cameras, RFID [22] [26] [29] [31] readers, factory-line sensors, fitness machines, and medical devices. Ubiquitous connectivity and the growth of sensors and intelligent systems have opened up a whole new storehouse of valuable information. Edge data can provide significant value to both the private and public sector as a source of enormous potential for gaining deeper, richer insight faster and more cost-effectively than in the past. In many cases, analysis of edge data can help organizations respond to events and solve problems that were previously out of reach.

3. Implications for Technology

For data to be analyzed where it resides, compute and storage capabilities must be local at the edge and in the cloud. This local infrastructure must address a set of unique challenges based on characteristics of the data and related issues.

- Sensed data is massive and streams 24-7.
- Data is noisy and dirty and requires preprocessing.
- Data has strong locality characteristics, meaning that the devices are operated and consumed locally.
- Data ownership, interoperability, security, and privacy are big issues.

How does this translate into a real-life example? Here's a transportation and public safety example.

- Road sensors may belong to different departments.
- Some cameras are owned by public security, while others belong to public transportation.
- Data is generated on private vehicles.

The issues: Can the data from these multiple systems be integrated and analyzed for meaningful insight? Who owns the data generated on private vehicles? Is the data secured?

These issues are well worth resolving. Multiple data stream scan unlock intrinsic correlations [4] [5] [16] that can have great significance overall. A recent study in a city in the People's Republic of China (PRC) shows that if you can detect morning wash time from the water supply subsystem, you can infer the morning rush hour; similarly, if you can detect when offices are powered down in the evening, you can infer the evening rush hour. Understanding these relationships can help cities better handle traffic at peak times as well as improve availability of water and electrical resources when they are most needed.

4. What's next?

Big data is a game changer and it's already here. While most of the momentum around big data today is around social media sources, I believe that realizing the promise of big data [1][2][21] analytics must include a way to harness the potential of big data from intelligent systems and sensors.

- Understand use cases and their implications. We must understand how existing disparate data sources can be evolved into a network of integrated, intelligent, connected systems.
- Define the usage model requirements for the analytics of edge data. The architecture must take advantage of big data distributed frameworks [24] [27] to move computation closer to where the data resides and support big data analytics at the edge via intelligent systems and local clouds.
- Enable the fast and secure delivery of aggregated data from edge analytics systems [27] [28] to other cloud and analytics platforms for further analysis.
- Address issues related to data ownership, interoperability, security, and privacy.

4.1. Take the Next Steps to Manage and Analyze Edge Data

Here's how you can get ready to take advantage of this fast moving area for your organization.

- Keep up-to-date with what's happening. For example Intel offers practical guidance to help you deploy big data environments more quickly and with lower risk.
- Explore business opportunities deriving from the analytics of edge data. Collaborate with the business to understand existing edge systems and the potential use for data. For more information

References

[1] http://www.webopedia.com/TERM/B/black berry_playbook.html
[2] White, Tom (2012). Hadoop: The Definitive Guide. O'Reilly Media p 3 ISBN 978-1-4493-3877-0
[3] MIKE2.0, Big Data Definition"
[4] Kusnetzky, Dan. "What is "Big Data?"". DNet Vance, Ashley (2010). "Start-Up Goes After Big Data With Hadoop Helper". New York Times Blog

[5] a b c d e f "Data, data everywhere". The Economist (25) 2010 Retrieved (9) 2012
[6] "E-Discovery Special Report: The Rising Tide of Nonlinear Review". Hudson Global Retrieved 2012.by Cat Casey and Alejandra Perez
[7] "What Technology-Assisted Electronic Discovery Teaches Us About The Role Of Humans In Technology — Re-Humanizing Technology-Assisted Review" Forbes. Retrieved 2012
[8] Francis, Matthew (2012). "Future telescope array drives development of Exabyte processing". Retrieved 2012
[9] "Community cleverness required". Nature 455 (7209): 1. (4) 2008. Doi: 10.1038/455001a.
[10] Sandia sees data management challenges spiral". HPC Projects 2009
[11] Reichman, O.J.; Jones, M.B.; Schildhauer, M.P. (2011)"Challenges and Opportunities of Open Data in Ecology" Science 331 (6018): 703–5. doi:10.1126/science.1197962.
[12] Hellerstein, Joe (2008). "Parallel Programming in the Age of Big Data" Gigaom Blog
[13] Segaran, Toby; Hammerbacher, Jeff (2009). Beautiful Data: The Stories behind Elegant Data Solutions. O'Reilly Media.p 257.ISBN 978-0-596-15711-1
[14] a b Hilbert &López 201.
[15] "IBM what is big data? Bringing big data to the enterprise" www.ibm.com. Retrieved 2013
[16] Oracle and FSN, "Mastering Big Data: CFO Strategies to Transform Insight into Opportunity", 2012
[17] Jacobs, A. (2009). "The Pathologies of Big Data" ACMQueue
[18] Magoulas, Roger; Lorica, Ben (2009). "Introduction to Big Data" Release 2.0 (Sebastopol CA: O'Reilly Media) (11).
[19] a b Snijders, C., Matzat, U., &Reips, U.-D. (2012). 'Big Data': Big gaps of knowledge in the field of Internet. International Journal of Internet Science, 7, 1-5. http://www.ijis.net/ijis7_1/ijis7_1_editorial.html
[20] Hogan, M. (2013). "Large Databases"
[21] Douglas, Laney. "3D Data Management: Controlling Data Volume, Velocity and Variety". Gartner Retrieved 2001
[22] Beyer, Mark. "Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data". Gartner Archived from the original on 10 July 2011.Retrieved 2011
[23] Richard Waters (2013). "Google search proves to be new word in stock market prediction". Financial Times Retrieved 2013
[24] David Leinweber (2013). "Big Data Gets Bigger: Now Google Trends Can Predict The Market". Forbes. Retrieved 2013
[25] Jason Palmer (2013). "Google searches predict market moves". BBC Retrieved 2013
[26] Kalil, Tom. "Big Data is a Big Deal". White House Retrieved 2012
[27] Executive Office of the President (2012). "Big Data across the Federal Government" White House Retrieved 2012
[28] "How big data analysis helped President Obama defeat Romney in 2012 Elections". Bosmol Social Media News 2013 Retrieved 2013
[29] Hoover, J. Nicholas. "Government's 10 Most Powerful Supercomputers" Information Week UBM Retrieved 2012
[30] Bamford, James. "The NSA Is Building the Country's Biggest Spy Center (Watch What You Say)". Wired Magazine Retrieved 2013
[31] Groundbreaking Ceremony Held for $1.2 Billion Utah Data Center" National Security Agency Central Security Service. Retrieved 2013
[32] Layton, Julia. "Amazon Technology" Money.howstuffworks.com Retrieved 2013.