Chapter 17
Big Data Architecture:
Storage and Computation

Siddhartha Duggirala
IIT Indore, India

ABSTRACT

With the unprecedented increase in data sources, the question of how to collect them efficiently, effectively, and elegantly, store them securely and safely, leverage those stocked, polished, and maintained data in a smarter manner so that industry experts can plan ahead, take informed decisions, and execute them in a knowledgeable fashion remains. This chapter clarifies several pertinent questions and related issues with the unprecedented increase in data sources.

INTRODUCTION

We live in the age of Data. Eric Schmidt famously said in 2010 that every day we create as much data as was created in total from beginning of written history through 2003. With the propulsion of mobile devices, sensors, search logs, online search, digital social lives we are generating about 2200 Petabytes of data every day (Kirkpatrick, R., 2013).

Google, Amazon, Facebook, Twitter, Foursquare, McDonalds and lots of other companies build their empires, enriched those empires using the data we generate (Kohavi, R., 2009).

- That being said, what is data? Data is a collection of facts, opinions and responses.
- Is Big Data (or even extreme data as some people like to call it) nothing but hyped version of normal data? Well, the major distinction comes from the 3 V's Volume (Petabytes per day), Variety (structured data like RDBMS, Unstructured like search logs, tweets, images, videos excreta), Velocity (real time capture) characterizing Big Data. While the traditional data mainly sit in RDBMS, Big Data otherwise the extreme data also encompass a different domain of data storage other than normal structured data.

DOI: 10.4018/978-1-4666-9840-6.ch017

Copyright © 2016, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

- Mere definition of Big Data would be "A massive volume of both structured and Unstructured data that is so large that it's difficult to process with traditional databases, software techniques" (Big Data: New frontiers of IT Management)
- Why do we have to store and analyze this data anyway? Simply there is a lot of potential in data which when observed at, analyst can create world class wonders. There has been a lot of research on this and these are the few reports you can go through? (Bryant, R.E., 2008; Manyika, Brown, 2011)

Since, we answered the questions "Why Big Data?", "what is Big Data?" let's answer most relevant basic question to us, How to store and leverage Big Data? Let's start answering the question by agreeing on the fact that Big-Data isn't just data growth, nor is it a single technology; rather, it's a set of processes and technologies that can crunch through substantial data set quickly to make complex, often real-time decisions. We will study technological and technical advancements that fueled Big data phenomenon, in the next section.

In 3rd section we will move on to Hadoop, Sector-Sphere and various other software frameworks enabling us to compute at Big data scale.

TECHNICAL AND TECHNOLOGICAL ADVANCEMENTS

There are a lot of ways to store and analyze data. Let's move on to technologies that enabled us to analyze data. And let's also look at various analyses that are predominantly used on Big data.

A/B Testing

A/B testing as the name sounds we have to decide which version A and B is better. To do this we experiment simultaneously. At the end we select the version which is more successful (Brain, 2012).

Associated Rule Learning

Set of techniques for discovering interesting patterns/relationships among variables in large databases.

Beowulf Cluster

The project started in mid-90. It was initially a cluster of 16 Dx4 connected by channel bounded Ethernet links. The cluster structure would be of parent and children kind of hierarchical structure. The client submits jobs to parent node, which in turn handovers the jobs and data to children node for processing and which in turn sends output to the parent node, parent node aggregates the output of children node do some further processing and gives out the final output to the client. Writing the programs for child node and parent node might get a little tricky.

28 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:
www.igi-global.com/chapter/big-data-architecture/150173

Related Content

Data Mining Using Fuzzy Decision Trees: An Exposition from a Study of Public Services Strategy in the USA
Malcolm J. Beynonand Martin Kitchener (2010). *Data Mining in Public and Private Sectors: Organizational and Government Applications (pp. 47-66).*
www.irma-international.org/chapter/data-mining-using-fuzzy-decision/44282

Semi-Supervised Sentiment Classification on E-Commerce Reviews Using Tripartite Graph and Clustering
Xin Lu, Donghong Gu, Haolan Zhang, Zhengxin Song, Qianhua Cai, Hongya Zhaoand Haiming Wu (2022). *International Journal of Data Warehousing and Mining (pp. 1-20).*
www.irma-international.org/article/semi-supervised-sentiment-classification-on-e-commerce-reviews-using-tripartite-graph-and-clustering/307904

Extending LINE for Network Embedding With Completely Imbalanced Labels
Zheng Wang, Qiao Wang, Tanjie Zhuand Xiaojun Ye (2020). *International Journal of Data Warehousing and Mining (pp. 20-36).*
www.irma-international.org/article/extending-line-for-network-embedding-with-completely-imbalanced-labels/256161

Discovering Similarity Across Heterogeneous Features: A Case Study of Clinico-Genomic Analysis
Vandana P. Janeja, Josephine M. Namayanja, Yelena Yesha, Anuja Kenchand Vasundhara Misal (2020). *International Journal of Data Warehousing and Mining (pp. 63-83).*
www.irma-international.org/article/discovering-similarity-across-heterogeneous-features/265257

Multi-Label Classification: An Overview
Grigorios Tsoumakasand Ioannis Katakis (2007). *International Journal of Data Warehousing and Mining (pp. 1-13).*
www.irma-international.org/article/multi-label-classification/1786