Landscape of Unified Big Data Platforms

Xiongpai Qin
Renmin University of China, China

Biao Qin
Renmin University of China, China

Cuiping Li
Renmin University of China, China

Hong Chen
Renmin University of China, China

Xiaoyong Du
Renmin University of China, China

Shan Wang
Renmin University of China, China

INTRODUCTION

Big data has become a hot topic in recent years (Mayer-Schonberger & Cukier, 2013). Big data is defined by WikiPedia (WikiPedia, 2013) as a term "for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications." The sources of big data include e-commerce, Internet of things, scientific experiments, and Web applications such as user generated contents etc. People have agreed on several *V*s to describe big data, including volume, velocity, variety, and veracity.

When processing huge volume of data of various types (structured data, unstructured data, and semi-structured data), at different stages (data in motion, data at rest, and archived data) to extract valuable information for decision making, people face several challenges, including capturing, cleaning, storage, transferring, searching, analysis, and visualization of the big data. The challenges sparkle the development of unified big data platforms. This chapter discusses the current competition landscape of unified big data platforms.

BACKGROUND

This section introduces the rising of a new parallel computing technology for big data - MapReduce, and compares MapReduce against RDBMS (relational database management systems).

Rising of MapReduce

MapReduce was introduced by Google in 2004 (Dean & Ghemawat, 2004) to process big volume of unstructured data. Now it has become a standard tool for big data processing and analytics. In industry, dozens of big data startups are launched, building their businesses around the MapReduce technology. They are Cloudera, HortonWorks, MapR, Karmasphere, DataMeer, Aster Data, Greenplum, Hadapt, and Platfora etc. In academia, MapReduce has aroused tides of research in parallel computing and database community.

The research has touched almost every aspect of MapReduce (Lee, Lee, Choi, Chung, & Moon, 2011; Sakr, Liu, & Fayoumi, 2013), including: (1) Storage layout, data placement, handling of data skew, index support, and data variety sup-

DOI: 10.4018/978-1-4666-5202-6.ch125

Copyright © 2014, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

port. (2) Extension of MapReduce for stream processing, incremental & continuous processing, iterative processing, leveraging of large memory of the cluster. (3) Optimization of joining, parallelization of complex analytical algorithms. (4) Schedule strategies for multi core CPU, GPU, heterogeneous environment, and cloud. (5) Easy to use interfaces for SQL, statistical algorithms, data mining & machine learning algorithms. (6) Energy saving, private and security guarantee. Due to space limitation, readers can refer to the two above mentioned references for more details. The references can be used as two hubs to recent research literatures.

Hadoop is an Apache project founded by Doug Cutting, and the Hadoop software stack is an open source implementation of the MapReduce technology. Throughout this paper, the two terms of MapReduce and Hadoop are used interchangeably. MapReduce is used when we express the general concept of MapReduce computing model (not specifically referring to Google's MapReduce platform), and Hadoop is used when some products of vendors, which are based on Hadoop, are introduced.

Traditional players in the database market also noticed the popularity of MapReduce. IBM moves quickly with its *Big Insights* Plan, which tries to integrate DB2, Hadoop, Netezza, and SPSS into a big data analytic platform. TeraData acquired Aster Data to obtain its experience of MapReduce as well as the analytic software package using MapReduce-style parallelization. EMC, formerly not as a database vendor, became a big player in the market overnight through acquiring Greenplum. Several venders who looked down on MapReduce before finally change their minds, Microsoft rejected MapReduce in 2009, and in 2012 it has closed the Dryad project (Foley, 2011) (a parallel computing framework similar to MapReduce) and warmly hugs Hadoop. Oracle despised MapReduce in early 2011, finally published its *Big Plan* which involved noSQL/Hadoop providing in late 2011.

Strengths and Weaknesses of RDBMS and MapReduce

Relational model has been extensively studied since 1970s. Various storage, indexing, optimization and execution techniques have been investigated and implemented to make RDBMS a mature tool for data processing in OLTP (online transaction processing) and OLAP (online analytics processing) applications. RDBMS has several advantages, including separation of physical storage and logical schema, high data consistency (ACID, i.e. atomicity, consistency, isolation, durability guarantee), high reliability, and high performance. In the age of big data, RDBMS encounters some difficulties. Firstly it does not scale well. RDBMS has not been deployed onto a cluster of more than 1000 nodes. Secondly RDBMS can not handle semi-structured data or unstructured data well, for example, it is difficult for RDBMS to handle graph data efficiently.

MapReduce is a general execution engine whose runtime system automatically parallelizes computation tasks across a large cluster of commodity servers, and handles failures. The user only needs to provide a map function (applied to all input rows of the dataset to produce an intermediate output) and a reduce function (aggregate intermediate results to produce the final result). MapReduce is designed to be highly scalable and highly fault tolerant to run on large clusters. The model is perfect to compute statistics from Petabytes (PB) of data. Furthermore, various data analytic algorithms have been migrated onto the MapReduce platform, including SQL query, machine learning, and data mining algorithms (Table 1).

Convergence of RDBMS and MapReduce

RDBMS is a mature technology. A data processing and analytic ecosystem (vendors, products, tools, services…) built around RDBMS has been there

10 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/landscape-of-unified-big-data-platforms/107333

Related Content

Scheduling of Extract, Transform, and Load (ETL) Procedures with Genetic Algorithm
Vedran Vrbaniand Damir Kalpi (2015). *International Journal of Business Analytics (pp. 33-46).*
www.irma-international.org/article/scheduling-of-extract-transform-and-load-etl-procedures-with-genetic-algorithm/126832

Future of Business Intelligence in Cloud Computing
Krishan Tuli (2021). *Impacts and Challenges of Cloud Business Intelligence (pp. 214-227).*
www.irma-international.org/chapter/future-of-business-intelligence-in-cloud-computing/269821

Recommendation and Sentiment Analysis Based on Consumer Review and Rating
Pin Ni, Yuming Liand Victor Chang (2020). *International Journal of Business Intelligence Research (pp. 11-27).*
www.irma-international.org/article/recommendation-and-sentiment-analysis-based-on-consumer-review-and-rating/258604

Social Network Analysis
Roberto Marmo (2014). *Encyclopedia of Business Analytics and Optimization (pp. 2221-2230).*
www.irma-international.org/chapter/social-network-analysis/107408

Overview of Predictive Modeling Approaches in Health Care Data Mining
Sunita Soni (2016). *Business Intelligence: Concepts, Methodologies, Tools, and Applications (pp. 73-95).*
www.irma-international.org/chapter/overview-of-predictive-modeling-approaches-in-health-care-data-mining/142612