Zarada Business Analytics

White Paper
June, 2011

Distributed Computing in Business Analytics

More the data, better is the analytical model

It is a thumb rule of statistics. The effectiveness of statistical modeling not only depends upon appropriate algorithms but also on high quality large data sets. So it's very important to have as much past data as possible to get better picture of future. But large data sets come with their own price. They require huge computational resources and the existing analytic products are not built for it. This white paper discusses why distributed systems are critical in next generation business analytic tools and explains a real-world architecture using a case study.

Business Analytics

According to Wikipedia, "Business analytics (BA) refers to the skills, technologies, applications and practices for continuous iterative exploration and investigation of past business performance to gain insight and drive business planning. Business analytics focuses on developing new insights and understanding of business performance based on data and statistical methods." In simple words, business analytics help the business to plan for the future using historical data.

Business Analytics is a critically important area from tools, solutions and implementation perspective, since more and more organizations are trying to formulate / validate their long term market strategies in addition to understanding customer behavior, cost relationships, customer satisfaction determinants etc by BA tools. BA tools have also evolved to expect minimum technology familiarity from BA users and have started providing recommendations as well as impacts in pure business terms. BA is no more a luxury to have, but it is the new necessity. It's a natural progression from traditional ERP systems. Importance of fact based decision making has gained significant mind share and it seems that there will be a large market for analytic tools in coming years.

Cloud computing and distributed systems

Cloud computing is in rage. Powerful servers and large computing power which was property of only handful of companies is now available to any enterprise through cloud computing. So, more and more enterprises are trying to leverage the cloud power to solve the complex problems which were impossible to solve few years back.

Cloud computing employs the distributed computing model where the applications will be running in multiple virtual machines rather than running in a dedicated physical server. This model is very effective since there is no more reliance on the physical infrastructure.

Business Analytics in Cloud era

Analyzing data in the cloud has gained tremendous momentum in last few years because of the attractive value proposition associated with the combination of business analytics and cloud infrastructure. Businesses are more and more convinced about the on-demand nature of cloud based resources as well as the cost-effectiveness, ease-of-use and ready availability of enterprise grade BA tools in cloud. Organizations can get started making useful analysis in a matter of days using cloud based tools which was unthinkable even two years back. Cloud based solutions truly provide a simple and cost-effective growth path and enables mid-size organizations to streamline their operational processes and promotes fact based decision making.

Drivers for cloud adoption

1. More and more enterprises want to run analytics on large amount of data.
2. Enterprises see great value in getting quickly started on analytics without much upfront investment as well as long term commitment to proprietary vendor technology.
3. Flexible, in terms of storage and computing capacity, enough to adapt to the changing business requirements.

Large data in Business Analytics

Traditional business analytical products developed for small data sets typically stored data in Microsoft excel or similar products. But as scope of business analytics is broadened, the amount of data used in analytics also increases. Enterprises, especially web based service providers, have huge amount of data to be processed for analytics. One of the good examples is the logs generated by user interactions in the websites.

Since there is huge number of users for the popular websites, enormous amounts of logs are generated for a given day. So, over the time, the data becomes so huge that it is very difficult to run the analytics using the traditional analytic tools. Also, large data may not always be structured which poses critical challenge to the traditional business analytic tools which are primarily designed to work with structured data.

Challenges of large data in Analytics

Performing analytics on large data set is not a straight forward problem. It inherently has many nuances which should be considered before developing solution for them.

1. Large data sets should be distributed across the machines rather than keeping them in a central place.
2. Processing of data sets should also be distributed across machines.
3. The statistical and data mining methods employed should support parallel processing.
4. Handling the latency of the distributed systems in the case of the interactive analytics.

Distributed computing in Business Analytics – A case study using Zarada

Zarada BA suite is next generation Business Analytics solution which provides analytic solution using data mining techniques applied on large data sets. It uses Apache Hadoop as the distributed computation and storage framework.

Apache Hadoop supports data-intensive distributed computing. It uses map-reduce algorithm along with a network file system and enables easy decomposition and distribution of data as well as computation load, thus effectively solving the large data challenges mentioned above in an elegant and cost effective way. Capacity of the cluster is increased simply by adding additional machines almost without any other manual intervention. With support for redundancy and simple management tools, it presents a viable platform to perform data mining activities on large data sets.

Large data distribution using Apache Hadoop

Apache Hadoop is built for large data sets. It supports a highly distributed file system called as HDFS (Hadoop distributed file system). This file system largely differs from the typical file systems found on operating systems like Windows, Linux etc. On traditional file systems block size, amount of data read in one clock cycle, is 4 kb where as in HDFS its 64MB.

HDFS primarily supports unstructured data stored in files and can effectively process multiple Terabytes of data since large data sets are supported from system level.

Zarada BA suite provides the Data Stream Orchestrator (DSO) product which allows users to distribute the data stored in database, csv files over the Hadoop distributed file system using a configuration-only task.

Distributed Processing in Apache Hadoop

Apache Hadoop employs map-reduce algorithm for distributed computing. Mappers are used to decompose a specific computation task whereas reducers are used to aggregate the results of multiple smaller computation tasks. Calculation logics are obviously provided by application developers.

Traditional distribution systems relied heavily on inter-process communication for distributed computing. Inter process communication is typically complex and, if left to users, becomes a major source of errors. Hadoop's inter process communication is built upon the Remote Procedure Call APIs of Java and uses locality of data to effectively distribute the processing so that very less data is transferred between the systems and throughput remains high. In essence, Hadoop takes out IPC headache from developer and provides a 'shared nothing' environment to simplify job creation and distribution.

Zarada BA Suite provides an easy-to-use, flexible workflow based system where user can configure different processing jobs that are distributed over the Hadoop. It hides the complexity of job dependency handling of Hadoop from user.

Choosing data mining methods for analysis

Though choosing the method of analysis is not directly dependent on the underlying distributed system, all methods or algorithms cannot be used for distributed computing.

Follows below are the traits of an algorithm to be qualified for distributed analysis:

1. It should support parallelism with respect to data. Since distributed systems are only effective when a computation intensive task can be divided into multiple smaller tasks and can be run separately in multiple machines. As for example, K-mean algorithm is parallel in nature.

2. If algorithm supports parallelism between different process steps, it can greatly improve the processing speed using distribution.

3. Algorithms with recursion and too many if / else or switch cases are not good candidate.

Zarada BA Suite provides a number of algorithms including Multiple Regression and Neural Networks to solve classification, prediction and optimization problems. Zarada created models uses parallel and distribution friendly algorithms to achieve high throughput using clusters with commodity grade hardware.

Interactive analytics and distributed system

Distributed systems are good for the large data processing, but they also add latency to the analytics. Since they are inherently slower than the non-distributed systems they are not good candidate for the interactive analytics. However, the problem can be resolved by adding intelligence in the system which adapts the underlying computing infrastructure depending upon the amount of data.

Zarada Business Analytics Suite provides an automated & intelligent

way wherein it can choose the type of processing depending upon the requirements at hand. When the data is small and need to be responded in real time, it runs analytics in memory. But when the data is large, it automatically leverages underlying Hadoop infrastructure for processing and distribution of large data.

Conclusion

Decisions drive organizations. A good and timely decision can enable an organization to launch a profitable product, to stop customer churn or to prevent registering a fraudulent customer. However, good decision needs awareness along multiple lines of facts. Some of these facts lie among divergent systems within organization and some outside the organization. Collecting, cleaning and analyzing large amount of data poses significant challenge both in terms of availability of in-house data analysts as well as large hardware requirements.

The need of the hour is to have an integrated & compact system which provides all the essential data analysis functions, namely collection, performance measurement & dissemination and predictive modeling and which doesn't require much of specialized statistical / mathematical knowledge to operate. If such a system can provide high scalability using small commodity grade hardware, then you probably have the best of our time.

Zarada Business Analytics is an easy-to-use and intuitive suite of products which enable you to handle large data and gain actionable insight using distributed computing clusters made of commodity grade hardware.

About Us:
Zinnia Systems is a software solutions company with a highly experienced team of technical and functional experts in the areas of business support systems and business analytics having years of experience in working with diverse customers.

Zinnia Systems. #2283, 14th 'A' Main, HAL 2nd Stage, Bangalore – 560 038.
Ph: +91-80-4110 0860/61, Fax: +91-80-4110 0862, Web:- www.zinniasystems.com
Reach us on contactus@zinniasystems.com