BIG DATA ANALYTICS – ARCHITECTURES & OPPORTUNITIES

Gus Verzosa

08/03/2013

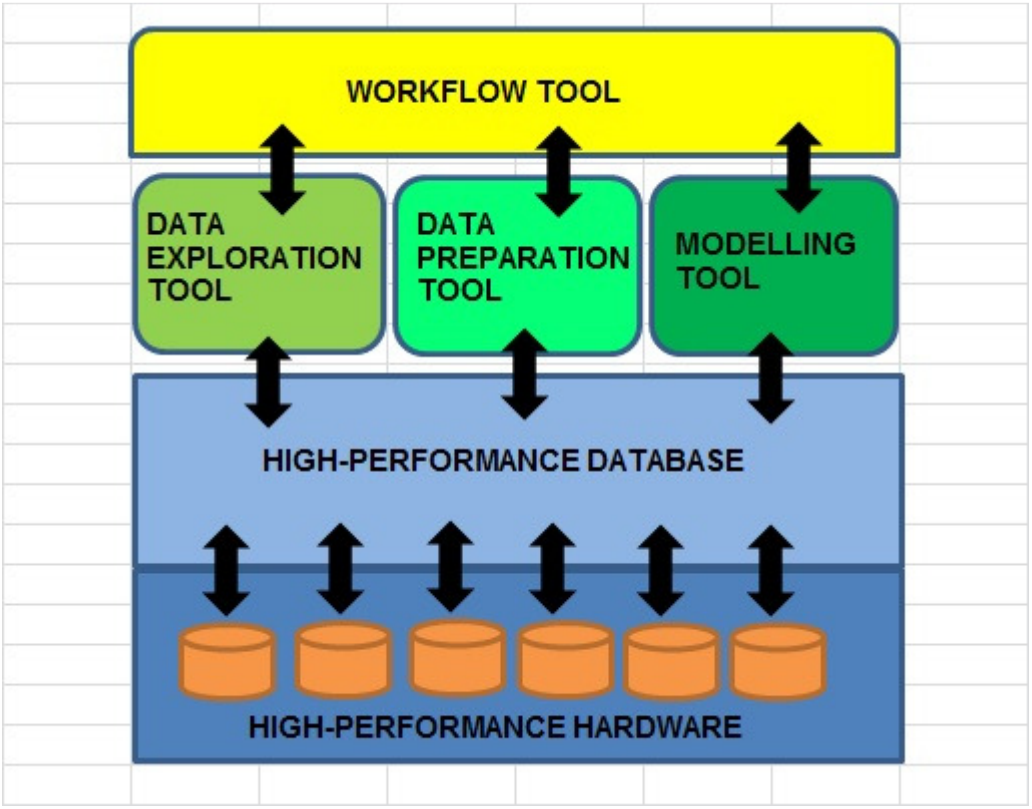
OBJECTIVES

The purpose of this paper is to provide an overview of:

1) The various emerging architectures for big data analytics.
2) The market opportunities for big data analytics.

ARCHITECTURES

The emerging big data analytics architectures utilize a combined front-end, back-end, and cross-architecture set of components. Please see the diagram below:



The back-end components are (collectively referred to as the analytic sandbox):

1) The high-performance hardware – this is typically a multi-node cluster, a large memory unit with back-up disk, or a combination of both. Parallel-processing hardware typically depends on multi-node clusters while in-memory hardware utilizes large memory units.
2) The high-performance database – this is a database tightly coupled to the high-performance hardware appliance. It normally accommodates structured (e.g. from relational data tables, from data warehouses), semi-structured (e.g. text data files, XML), quasi-structured (e.g. clickstream), and unstructured data (text documents, PDF, images, video).

The front-end components are:

1) The data exploration tool – for the Discovery phase of the Data Analytics Lifecycle.
2) The data preparation tool – for the Data Preparation phase of the Data Analytics Lifecycle.
3) The modelling tool – for the Model Planning and Model Building phases of the Data Analytics Lifecycle.
4) The workflow tool – for organizing the functions into a structured process during the Operationalization phase of the Data Analytics Lifecycle.

The cross-architecture component, which moves between the front and back is the In-database analytics component – comprising procedures that run within the database itself that speed up the analysis of large volumes of data.

Both the in-database analytics and the back-end high performance database and hardware (utilizing parallel processing, in-memory processing, or a combination of both) facilitate the capability to analyse with a high level of complexity large volumes of various types/formats of data quickly. This capability goes beyond traditional BI-EDW's and is required by big data analytics.

The architectural components above have been implemented by key vendors as follows:

VENDOR	SAS	IBM	SAP	EMC	TERADATA	OPEN SOURCE	ORACLE
COMPONENT							
HIGH-PERFORMANCE HARDWARE	Open	Netezza	HANA	EMC	Teradata	Open	Exadata
HIGH-PERFORMANCE DATABASE	Open	Netezza	HANA	Greenplum	Teradata	Hadoop	Exadata
DATA EXPLORATION TOOLS	SAS BI, EG, Analytics, etc.	Intelligent Miner, Cognos	Advanced Analytics	R, SAS	SAS	R, RapidMiner	OBIEE
DATA PREPARATION TOOLS	SAS DI, Dataflux	DataStage/QualityStage	BOBJ Data Services	R, SAS	SAS	R, RapidMiner	OBIEE
MODELLING TOOLS	SAS Enterprise Miner, Analytics	Intelligent Miner	Predictive Analytics	R, SAS	SAS	R, RapidMiner	OBIEE
WORKFLOW TOOLS	SAS EG (Enterprise Guide)	Lombardi, DS/QS	SAP CAF, BPM			RapidMIner	OBIEE

(continued next page)

OPPORTUNITIES

The big data opportunities in the various industries are as follows:

INDUSTRY	OPPORTUNITY
Public Sector	
DIAC	Analysis and prediction of contribution of various migrant groups to population growth, education output, and workforce distribution as input to immigration policy.
DEEWR	Correlation and prediction of education system output population versus industry workforce demand to support programs to minimize the skills shortage through aptitude optimization and industry matching.
DHS	Resource usage analytics and correlation using demographic, employment, and financial data to maximize impact of resource deployment through targetted programs.
AFP	Crime analytics and prediction to lower crime rates through preventive education programs.
ASIO	Terrorism analytics and prediction to prevent attacks through preventive identification and countermeasures.
ATO	Revenue analytics and correlation using demographic, employment, and financial data to maximize collection and minimize fraud through targetted education programs.
Airservices Australia	Incident analytics and data mining to identify environmental, human, and engineering factors that impact frequency and severity of incidents.
AMSA	Incident analytics and data mining to identify environmental, human, and engineering factors that impact frequency and severity of incidents.
Infrastructure	Transportation (including traffic) and infrastructure (including usage) analytics to support optimal engineering and operations research.
General	Workforce analytics to assess and predict resource and skill levels and distribution in order to support effective and efficient delivery of programs and services.
Defence	Terrorism analytics and prediction to prevent attacks through preventive identification and countermeasures.
	Logistic analytics to optrimize supply chain effectiveness and speed based on Velocity Management principles.
Manufacturing	Factory Physucs analytics and data mining to predict and minimize production botllenecks and optimize inventory levels.
	Equipment analytics to predict and minimize downtime through preventive maintenance.
Distribution	Logistic analytics to optrimize supply chain effectiveness and speed based on SCOR principles and benchmarks.

(continued next page)

INDUSTRY	OPPORTUNITY
Retail	Demand and sales analytics to predict product demand and product-mix by customer segment and maximize sales.
	Customer analytics to segment and maximize customer lifetime value through targetted products and services.
Telecommunications	Churn analytics to predict and minimize churn.
	Fraud analytics to analyze and identify fraud.
	Usage and traffic analytics to support dynamic network reconfiguration for optimal bandwidth allocation and distribution.
	Equipment analytics to predict and minimize downtime through preventive maintenance.
Power	Fraud analytics to analyze and identify fraud.
	Utilization analytics to support dynamic power network reconfiguration for optimal power allocation and distribution.
	Equipment analytics to predict and minimize downtime through preventive maintenance.
Financial services	Churn analytics to predict and minimize churn.
	Fraud analytics to analyze and identify fraud.
	Asset-liability matching and prediction for optimal investment of resources.
	Customer analytics to segment and maximize customer lifetime value through targetted products and services.
	Risk and underwriting analytics and prediction to minimize losses.
Healthcare	Patient and clinical data to support evidenced-based treatment programs to improve patient health and care.
	Billing, inventory, payroll, and other cost analytics to streamline the healthcare supply chain and minimize the cost of healthcare to patients.
Mining	Equipment analytics to predict and minimize downtime through preventive maintenance.
	Logistic analytics to optrimize supply chain effectiveness and speed based on SCOR principles and benchmarks.
	Factory Physucs analytics and data mining to predict and minimize production botllenecks and optimize inventory levels.
	Geological data analytivcs to predict, locate, and assess resources.