Big Data Analytics
2015
by H. Michael Covert and Victoria Loewengart
Agenda

• Big Data Analytics market Sector Analysis
• Real time analytics
• Frameworks and Fabrics
• Systems Integration – living in the enterprise
• Pervasive analytics
  – From Predictive to Prescriptive analytics
• Looking ahead
The Big Data Analytics Market

• It is growing at a tremendous rate of 30% CAGR, and is expected to reach $114 B by 2018. Growth of this magnitude drives massive change.
Real Time Analytics

• **Major performance improvements** - Big data analytics moved well beyond their batch roots and now deliver real-time performance.
  
  • The emergence of in-memory computing has taken hold, and a new paradigm, the Lambda Architecture, has emerged as the primary means of delivering real-time analytics to the enterprise.

Nathan Marz
Big Data Analytics In the Enterprise

• **Significant advances in systems integration**
  – Vendors and open source initiatives have now more tightly integrated Big Data repositories into the enterprise.
  – The data lake concept has taken hold, yet effective usage requires analytical and data manipulation skills of the Data Scientist.
Frameworks and Fabrics

• Big data analytics systems have been notoriously difficult to program.
• The underlying fabric is changing very rapidly.
• New frameworks have now emerged that make software development easier, and also provide insulation from rapidly evolving infrastructure changes.
• These frameworks continue to advance machine learning and deep learning paradigms allowing sophisticated delivery of more complex forms of analytics.
Advanced Analytics

• Predictive analytics has evolved to prescriptive analytics.
• These have emerged as a desired end state in big data.
• Systems today use machine learning, graph theory, natural language processing, deep learning, and business rules to:
  – Discover trends and relationships
  – Predict future behavior or events
  – Prescribe actions if the prediction is adverse or not optimal
Market Sector Analysis

- Market sector growth has continued
  - Professional services
  - Applications
  - Storage
  - NoSQL
Market Sector Analysis

Figure 7: Big Data Analytics Market Size by Business Category

Source: Heavy Reading

COMPANIES ARE SPENDING BIG ON BIG DATA

In 2015
- Financial Services: $6.4B
- Software/Internet: $2.8B
- Government: $2.8B
- Comms & Media: $1.2B
- Energy/Utilities: $800M

Annual Growth to 2020
- Financial Services: 22%
- Software/Internet: 26%
- Government: 22%
- Comms & Media: 40%
- Energy/Utilities: 54%
Market Sector Analysis

Figure 1: Potential payback of big data initiatives

Source: Gartner, 2013.
The new reality is that opportunities abound and many are no longer “overeager.”
Real Time Analytics

• Big data analytics has now moved beyond batch
• Some have been “retrofitted”
  – HIVE and Pig
• New technology has also arrived
  – YARN
  – Tez
  – Spark
  – Storm
  – Impala
• A reference architecture stack has now been defined
Real Time Analytics

• **YARN** (Yet Another Resource Negotiator) is the basis for Hadoop version 2
  – Hadoop version 1 developed scalability problems due to its server architecture
    • The JobTracker had too much work to do
      – Allocating cluster resources
      – Managing running jobs
    • So, it was split into smaller components that each perform specific services
  
• These servers can now remain “persistent” and this forms the basis of real time processing
Real Time Analytics

• YARN has now become the new substrate (fabric)
Real Time Analytics

- YARN has now become the new substrate (fabric)
Real Time Analytics

- YARN has now become the new substrate (fabric)
Real Time Analytics

- **Tez** – a new architecture designed to create more flexible execution graphs and to minimize data exchanges (and writes to disk)
- It can be retrofitted into some existing systems
Real Time Analytics

• **Spark** is a new architecture that is designed for real-time analytics
  – It also uses a persistent server structure
  – It utilizes in-memory datasets called Resilient Distributed Datasets (RDD)
  – It provides a graph database component called GraphX
  – It provides a Machine Learning component called MLLib
Real Time Analytics

- **Storm** is a Complex Event Processing (CEP) system designed to handle high velocity streamed data
  - It was invented by Twitter but is Apache open source
  - It gets its data from “spouts” and processes its in “bolts”
  - Each bolt is a programmable unit which can transform, store, apply rules or machine learning in order to detect patterns in real-time.
  - Storm runs on top of YARN
Real Time Analytics
Development Frameworks

- CONCURRENT CASCADING is a very powerful framework
  - It makes MapReduce easier
  - It is dynamically controllable (unlike Pig, Hive, MapReduce, etc.)
  - It makes code reuse easier
  - It is portable across fabrics
    - Hadoop MapReduce
    - Hadoop YARN
    - Tez
    - Spark and Storm (soon)
Development Frameworks

• **Concurrent Cascading** is also the basis for a product “family” that uses its DSL
  – Lingual – a SQL query that produces big data workflows
  – **Pattern** – a Machine Learning interface that uses PMML
    • Note that this alone facilitates integration with SAS, R, and other ML development frameworks and languages
  – Other languages are supported
    • Cascalog – Cascading on Clojure
    • Scalding – Cascading on Scala
Systems Integration

• Big Data Analytics has really become part of the enterprise architecture, and now delivers a full stack
  – ETL tools now integrate nicely allowing access to HDFS
    • Pentaho, Talend, other commercial apps. And many now come with integrated BI and analytics components
  – Systems management is now solid
    • This was essential since management of a cluster is “tricky”
    • Cluster members (nodes) fail, and also new ones are added for expanded capacity
    • All nodes must run identical software stacks (OS, Java, Hadoop et al)
    • Cloudera Manager, Ambari, Mesos, and other Web based tools manage all of this
  – Security concerns are (mostly) addressed
    • Authentication, Data in Motion, Data at Rest
    • Accumulo and NIST/FIPS-140
The Rise of DevOps

- Tools like Concurrent Driven now provide deep analysis of applications, viewable down to the fabric
Systems Integration

• Big Data enterprise architecture has taken on new meaning
  – As we have seen, Hadoop is no longer the only option (i.e. with Spark, Tez, YARN, etc.)
  – As a result, new operating frameworks have been created – The Distributed System Kernel model

• Manages the cluster
• Allows multiple development frameworks to run on the same hardware cluster
Systems Integration – The Data Lake

Future Corporate Data Architecture

FAST DATA

- Fast Operational Database
- Ingest / Interactive
  - Real-Time Analytics
  - Decisioning
  - Export

BIG DATA

- Data Lake (HDFS, etc)
- BI Reporting
  - SQL on Hadoop
  - Exploratory Analytics
  - Map Reduce

ETL

- CRM
- ERP
- Etc

Enterprise Apps
Systems Integration – The Data Lake

Looks a little like a DW/CIF model
Systems Integration – The Data Lake

DATA

VALUE

PB's
PB
TB's

Data Lake

Risk Management
E.g., Fraud Reduction

Customer Intimacy
E.g., 360 Degree View of the Customer

Operational Excellence
E.g., Network Maintenance

New Business
E.g., Data as a Product

DATA LAKE
An architectural shift in the data center that uses Hadoop to deliver deep insight across a large, broad, diverse set of data at efficient scale
Pervasive Analytics

• Predictive Analytics
  – It has become pervasive.
  – Gartner Says by 2016, 70 Percent of the Most Profitable Companies Will Manage Their Business Processes Using Real-Time Predictive Analytics or Extreme Collaboration
  – Some of these solutions, IDC says, “will begin to replace or significantly impact knowledge worker roles.”
  – Frey & Osborne of the University of Oxford state that up to 66% of the U.S. workforce has a medium to high risk of being displaced by technology in the next 10-20 years. The time previous spent “managing machines” will be re-purposed.

Thornton May - January 2015

• To augment or replace (or maybe a little bit of both)?
From Predictive to Prescriptive Analytics

• And now, the bar has been raised. We need to know more than what. We need to know why and how.
  – Simple correlation moves to causation!
Text Analytics and Natural Language Processing

• **Text analytics** is another “old but now new again” trend
  – Reading and understanding text
  – Heavily reliant on machine learning
  – Areas of focus:
    • Sentiment analysis
    • Extraction of “named entities”
      – Connecting named entities through references, actions, etc.
    • Grouping documents with similar characteristics
    • Assigning documents to “topics”
    • Similarity
Text Analytics and Natural Language Processing

- What does Big Data have to do with Text Analytics and Natural Language Processing?
  - There are now a million words in the English language and about 3.4 billion combinations!
    - This is clearly a Big Data problem
  - Refining and improving language recognition benefits from massive amounts of data
    - Original datasets were relatively small – 42,000 sentences
    - Newer datasets are huge – 1,000,000 sentences and more
    - Language processing is a classic “long tail” distribution
    - Google “billion word” project
More and more need exists – free text data is exploding due to social media, voice recognition, and other “automated” systems.

The belief is therefore that Big Data will provide better capabilities for understanding language.

Machine learning has become key. Rule based NLP is still used, but most new science is statistical.
Text Analytics and Natural Language Processing

- Need for NLP skills is increasing
  - Jobs are showing up everywhere – 51% increase in job postings in the US alone
- From a job search perspective, problem domains are expanding
- It has moved out of academia into industry
- Both academic and corporate research is producing new results
  - Renewed interest in ontologies
  - Machine learning approaches are breathing new life in a “very difficult to do” subject.
- Even Bill Inmon is talking about it!
Text Analytics and Natural Language Processing

- Text analytics has become very important to:
  - Health Care – clinical notes, patient charts, diagnosis and procedures
  - Marketing – churn analysis, sentiment
  - Customer service – routing, sentiment
  - Political – sentiment, topic importance
  - Intelligence and Law Enforcement – intent, state of mind, sentiment
  - Intellectual Property and Competitive Intelligence
  - Legal, regulatory, and compliance
Big Data Analytics - What is next?

• We expect big data analytics growth to continue
  – Corporate adoption is becoming easier because of vendor support
  – We are very, very short on skill sets! Gartner says 183,000+ by 2016

  Big data -> Machine learning -> Deep learning -> Prescriptive Analytics

• We expect innovation and change to continue at a breakneck speed
  – As always, some danger exists in choosing a sustainable solution. The new paradigm often is not concerned with backward compatibility!
  – Look for “reference architectures” as a guide
  – Follow vendor integration strategies
  – Insulate yourself with frameworks that will give your development efforts longevity
Big Data Analytics – The question is...

- How do I get involved with it?
- What skills do I most need?
- How will it help my career?
Questions and Answers

Michael.Covert@AnalyticsInside.us
Victoria.Loewengart@AnalyticsInside.us

http://www.AnalyticsInside.us