Holistic Approach to Big Data #1: Introduction to Big Data

1

In these videos we will present a holistic approach to BigData, taking both a top-down and a bottom-up approach to questions such as

- What is Big Data?
- How do we tackle Big Data?
- Why are we interested in it?
- What is a Big Data platform?

In addition we will venture briefly into futures and the role of cloud technology in relation to BigData.

2

This slide lists the main topics that we will cover:

- The State of Big Data Adoption
- Big Data – A Holistic Approach
- Five High-value Big Data Use Cases
- Technical Details of Key Big Data Components
- The Future of Big Data and Cloud, and
- Available Resources

3

The first of these is the current state of BigData adoption.

For this we will draw upon a report

Analytics: The real-world use of big data in financial services

The IBM Institute for Business Value partnered with the Saïd [pronounced “sayeed”] Business School at the University of Oxford to conduct the 2012 Big Data @ Work Survey, the basis for the research study, surveying 1144 business and IT professionals in 95 countries, including 124 respondents from the banking and financial services markets, or 11 percent of the global respondent pool.

Participants were given a survey where they could provide multiple answers to “Big Data Sources” and “Analytics Capabilities” headings for where Big Data is being used.

- 88% of respondents (global) said they use transactions as their big data source
- 73% of respondents (global) said they use log data as their big data source

Because the respondents could choose more than one answer, these values overlap and thus do not sum to 100%

4

The Big Data adoption process goes through a number of phases:

- Education
- Exploration
- Engagement, and then
- Execution of a Strategy

Underneath each block on the chart are percentages of global sectors and the percentages Banking/Financial Management respondents that are at each stage.

5

In this chart, the left column shows sources for BigData, and the right column what to do with it.

Transaction data is EXISTING DATA that customers say is the main source of big data…and they want to query/report on it. In an RDBMS used for transactions, every time you update/delete/insert and select (queries), this information is logged. The business user can analyze this data for business purposes, not just for recovery purposes: thus, what is this information used for? The right column in this case shows "query and reporting."

The second entry on the left is "log data." Walmart is probably the largest company using RFIDs (radio frequency ID technology) to allow it to track everything it has to deal with around the world. This is sensor data that can be helpful to analyze use and inefficiencies. Thus, it is faster to get supply from the North of China than South of Vietnam?

Social Media (FB, Twitter, etc) is #5. Lots of buzz around it, yet not that important in this study. It's not that companies are not working on getting FB and Twitter data, but the emphasis is mainly on their existing transaction data (ranked #1) and log data (or machine data) ranked #2.

Mainframe machine data is considered to be trusted data, and based on this data from the mainframe (MF) is center piece for a BigData strategy.

Also, using "Still images/Video" and "Audio" are not yet important generally as a BigData resource. But, one should note, that clinical data, surveillance data, mobile cameras, etc. are actually around 70-80% unstructured data. Currently you will see "Analytics Capabilities" (in the 2nd column) of video analytics and voice analytics on the low end.

6

Big Data skills are in high demand, especially Java language and experience with Hadoop distributions.

The new profession is Data Scientist. The data scientist represents an evolution from the business or data analyst roles. The formal training is similar, with a solid foundation typically in computer science and applications, modeling, statistics, analytics, and math. What sets the data scientist apart is strong business acumen, coupled with the ability to communicate findings to both business and IT leaders in a way that can influence how an organization approaches a business challenge. Good data scientists will not just address business problems; they will pick the *right* problems that have the most value to the organization.

The data scientist role has been described as “part analyst, part artist.” It has been said that “a data scientist is somebody who is inquisitive, who can stare at data and spot trends. It’s almost like a Renaissance individual who really wants to learn and bring change to an organization."

Data scientists are inquisitive: exploring, asking questions, doing “what if” analysis, questioning existing assumptions and processes. Armed with data and analytical results, a top-tier data scientist will then communicate informed conclusions and recommendations to an organization’s leadership structure.

7

Big Data has become a business issue, or at least an issue that business people are aware of. Look at the coverage its getting in the business press. From the Wall Street Journal: “Companies are being inundated with data” to the Financial Times article: “Increasingly businesses are applying analytics to social media such as Facebook and Twitter,” on to Forbes: “big data has arrived at Seton Health Care Family.”

Why is Big Data getting this type of coverage? Because it has the potential to profoundly affect the way we do business.

The quote on CNBC really exemplifies this “Data is the new Oil.”

Data is a natural resource that is growing bigger. Like any resource, it is difficult to extract. It comes in many types – or a huge variety. It is also difficult to refine, or analyze. Many organizations do not even tap into this natural resource – they ignore data, or they use it for just one purpose. This is largely because it is difficult to structure and restructure for different purposes. But some organizations have cracked the code, and they have figured out how to process and analyze data available to them, and they are utilizing it to achieve breakthrough outcomes.

8

If data is a natural resource, what is your company doing to capitalize on it?

Here, graphically, we can see some of the sources of Big Data.

All forms of data are represented and all are processed by different types of analytics to yield insight to the business organization.

9

Sensors are one of the biggest contributors of Big Data, enabling new applications across industries.

Examples include

- Telemetrics for auto insurance and vehicle monitoring & service
- Smart metering for energy & utilities organizations
- Inventory management and asset tracking in the retail and manufacturing sectors
- Fleet management in logistics and transportation organizations

Applications that rely on sensor-generated data have unique Big Data requirements to efficiently collect, store, and analyze the data to take advantage of its value.

Web log and database log data are major producers of machine generated data.

Let’s look at some other examples of sensor and machine generated data to expand our perspective.

10

There are 4 engines on the Airbus A380, the double-deck, largest commercial airplane in the world. Each A380 engine generates 1 PB of data on a flight, for example, from London (LHR) to Singapore (SIN).

11

The Large Hadron Collider (LHC), when operational last year, generated 1 GB of data every second

12

A new Radio Telescope (called the Square Kilometer Array) is currently being built in the southern hemisphere. The SKA will produce 20,000 PB / day in the year 2020 (*compared with the current internet volume of 300 PB / day*).

This is predicted to grow by a factor of ten when fully operational in 2028. The various data centers to process it will each handle 100 PB / day.

—-

You have now completed this video. In the next video we will look at Big Data as a platform and not just a collection of disjoint software packages. We will take a holistic approach to this Big Data platform.