Holistic Approach to Big Data #2: Big Data as a Platform

1

We will use a holistic approach to understand the nature and requirements of a big data platform, but first we will ask what is big data specifically, how is it studied, and what do we want to achieve by working with big data.

Let's start with two client examples from IBM customers using big data.

2

The first client is KTH, Sweden's Royal Institute of Technology (Sweden's leading technical university), where they analyze real-time data streams to predict traffic patterns

The goal for this project at KTH is to combine traffic data from urban congestion with other data to take the ability to predict and improve traffic flow to the next level. Today, the system collects a great deal of useful data, but researchers were looking for a way to gain more insight from it. With a smarter solution, researchers could help the population of a metropolitan area decide, for example, the best time and method to travel from one place in the city to another, or when to leave in order to catch a flight at the airport.

Researchers were alerted to the capabilities of IBM InfoSphere Streams software and adopted it as part of their technology platform. This led to a collaboration where IBM employees would learn more about traffic research and the researchers at KTH would get access to software which would enable them to concentrate on their research rather than on programming.

Data is gathered from a number of different sources, including cameras at entries and exits to the city, GPS data from taxis and trucks, weather information, public transport systems, and more.

All the information is processed by Streams software, which can handle all kinds of data (both structured and unstructured. Based on previous experience and the data collected from instrumented measuring points, this intelligent system can make predictions on how long it should take to travel from point A to point B, give advice on alternative routes and alternative transportion systems, and improve traffic flow in the metropolitan area.

This approach enables researchers at KTH to analyze large volumes of streaming data in real-time and can help transform the vision of SmarterTraffic into a real solution. When the solution is fully implemented, it will lead to smarter, more efficient, and more environmentally friendly traffic in the metropolitan area. In addition, the solution can be replicated in other areas worldwide. It also provides new insight into the mechanisms that affect a complex traffic system.

What makes it smarter?:

- **Intelligent:** Handles large data streams and data other than pure traffic-related data (such as weather) to enable researchers to analyze and predict traffic faster and more accurately than ever before.
- **Instrumented:** Gathers data from a number of different sources, including cameras at entries and exits to the city, GPS data from taxis and trucks, weather information, public transportion systems, etc.
- **Interconnected:** Enters gathered data into the InfoSphere Streams software, which can handle all types of data, both structured and unstructured.

3

In the second example, IBM has been collaborating with the University of Southern California (USC) Annenberg Innovation Lab (AIL) (part of the Journalism school) to explore how technology can be used by organizations from news outlets and journalists, to movie studios and retailers, in order to better understand, respond, and predict public sentiment.

One of the on-going projects is to do real-time sentiment analysis on Twitter data related to the Republican primary debates and Presidential debates.

The challenges include:

- How to process, filter, and analyze millions of public Twitter messages
- How to determine the sentiment of each message
- Predict debate winners and changes in candidate popularity

How does the system do its work?

- The Streams application receives *all* the tweets related to the Republican primary through a direct connection to the Twitter firehose (100% coverage) vs the Twitter public API (only 1% coverage)
- Twitter-specific NLP preprocessing (NLP = natural language processing) is performed on each tweet to analyze its sentiment and aggregate the data for each candidate. The output is then displayed on a web-based ise omterface (UI).
- The data is archived to be able to rewind or fast-forward to a particular time period

Natural language processing (NLP) is based upon work done at the University of Pennsylvania (the Penn Treebank Project), Stamford University, Carnegie Mellon University, and others.

The benefits of the approach provide

1) Real-time public feedback on candidates' answers to debate questions
2) Early prediction of debate winner(s) based on public opinion
3) Independent assessment of who won the debate based on public opinion versus just the opinion of political analysts

[References:

- Link to USC political debate video: http://www.annenberglab.com/eventdetail/80
- Live Real-time Twitter Sentiment Analysis: http://politics.twittersentiment.org/streams]

4

We should ask what *is* big data and *how* does one go about using it?

- **What?** Big Data is about deriving new insight from previously untouched data and integrating that insight into your business operations — data warehouses, business processes, and applications
- **How?** Big data is about the application of new tools to *do* MORE analytics on MORE data for MORE people

5

We can also take note of the four characteristics of Big Data — often called V cubed (V4).

Volume, Velocity, Variety, Veracity.

These are four characteristics, but they do not define Big Data. Big Data was defined by the *What* and the *How* on the previous slide.

6

The final question is why do you want to deal with more data?

— Because you want *new insight*

This new insight is not only for top level executives

— This new insight will be used to get people throughout the enterprise to run the business better and to provide better service to customers

7

We should also note that Big Data is **not just Hadoop**.

There are many examples where Hadoop is not entirely applicable: Cyber security, stock market, traffic control, sensor information, and monitoring trends in social media, amongst others.

What if your company has many silos of information? It can be difficult to move these silos of information to an Hadoop Distributed File System (HDFS) where Hadoop stores its data/

We also need to ask about data governance here. What is data governance? Can we trust the source of this data?

8

You may have heard of the term **NoSQL**, which essentially means "**not only SQL**" — since there is data which must be processed by means other than just the SQL Quary Language.

Perhaps we should say that Big Data is "not only Hadoop" to round things out.

Big Data is best thought of as a PLATFORM rather than a specific set of software.

That is why we used the term earlier: **Not only Hadoop**, or **Not just Hadoop**.

9

What follows in this video are components that can be used to Put together a Big Data platform.

10

Data warehouses are part of a big data platform that you may already be familiar with. They deliver deep insight with advanced in-database analytics & operational analytics. Data warehouses provide online analytic processing (or OLAP).

11

Stream computing is used to analyze streaming data and large data bursts for real-time insights. This provided real-time analytic processing (we wish to instroduce you to the term, RTAP).

12

And, of course, Hadoop provides cost-effectively analysis of petabytes of unstructured and structured data.

13

All of the data processed needs to be governed for data quality as part of the management of information lifecycle.

14

In addition, **Accelerators** are software libraries that can be used to jumpstart the development of applications by providing toolkits that will have the operators, functions, and connectors needed for specific types of data or application areas. Examples are:

- **Analytic Accelerators** — text analytics, geospatial, time-series, data mining
- **Application Accelerators** — financial services, machine data, social data, Telco event data
- **Industry Models —** comprehensive data models based on deep expertise and industry best practices

15

The next layer of a complete big data platform provides the tools needed to discover, understand, search, and navigate federated sources of big data.

16

—-

You have now completed this video. In the next video we will continue to look at Big Data as a platform and we will add a discussion of **data governance** for this platform..