Amreeta Choudhury

Lecture Notes 06/07/2017 Module 3

Class: 8:30-10:00 pm EST

Announcements:

Discussion forum post due on June 10th by 11:59 pm EST. Work on first draft by June 11th by 11:59 pm

Lecture Notes:

Article 1: The best data scientists get out and talk to people

https://hbr.org/2017/01/the-best-data-scientists-get-out-and-talk-to-people

Session Objectives:

1. Describe the drivers, motivations, practices, and success criteria associated with a typical organization's tech unit
2. Identify tech-specific vocabulary
3. Analyze typical challenges and frustration associated with working with business units
4. Data Characteristics
5. Modern Database Architecture
6. Introduction of the Data Lake

We discussed tech-initiated projects. There are 3 types of roles/people. Visionaries usually have a hit idea after every couple of ideas. Shapers take that idea then hand them over to the operations team who have to run them each day. There needs to be a natural flow from top to bottom with different evaluation criteria that needs to be considered.

You need to make sure that the tech centric evaluation and the business success criteria are speaking the same language. A tech evaluation may not be business related. It is not always easy to achieve both. For example, it is important to see how fast queries run but businesses may not care. We can't only use this as the success criteria.

Discussion of Project Management Methodologies:

Many of the slides were taken directly from the readings. See below for more links and information.

Agile was created in order to bridge the gap between tech and business. They wanted to elevate the programming to a profession and make it easier. They went into this expecting to have communication issues.

Check out blog.udemy.com/agile-vs-waterfall

Extreme Programming: http://www.extremeprogramming.org/rules.html

https://www.mountaingoatsoftware.com/blog/differences-between-scrum-and-extreme-programming

https://www.atlassian.com/agile/scrum

Perceptions of Agile

http://www.forbes.com/sites/stevedenning/2012/04/17/the-case-against-agile-ten-perennial-management-objections/#46bd26a6715e

http://www.cio.com/article/2385322/agile-development/why-agile-isn-t-working--bringing-common-sense-to-agile-principles.html

How to choose which method: Complexity of the problem, team experience, stability of requirements, customer involvement, delivery speed and quality, uniqueness and high regulatory requirements

Source: Cadle, J. (2014). *Developing information systems: Practical guidance for IT professionals*. London: BCS, The Chartered Institute for IT.

Good pages to consider for business draft:

Page 50 of Developing information systems book

Investigations: Requirements engineering: Pages 65+ (interviews, workshops, focus groups, shadowing, prototypes, to quantitative; surveys, document analysis, activity sampling), Contents of a requirements document page 70

Organization Roles and their personas:

Waterfall Key Roles

- **Developer**
- **Tester**
- **Business Analyst**
- **Project Manager**

Scrum Key Roles

- **The product owner**
- **The scrum master**
- **The scrum team**

How to get started in a tech project:

1. Proof of Technology
2. Proof of Concept
3. Pilot Project

Types of Tech Projects:

Line of business

Shared Services

Enterprise

Things to think about:

Data Source

Data

Logic

Presentation

Evolution of Data flow:

It started out internal within a company such as a transaction system with mostly structured data.

We moved to the ETL Process and the data model where the data was known.

Then came data extraction, reporting and business intelligence

Lastly, Presentation has come a long way in terms of UI, UX

ETL

- **Extract** is the process of *reading data* from a database.
- **Transform** is the process of *converting the extracted data* from its previous form into the form it needs to be in so that it can be placed into another database. Transformation occurs by using rules or lookup tables or by combining the data with other data.
- **Load** is the process of *writing the data* into the target database.

ETL is used to migrate data from one database to another, to form data marts and data warehouses and also to convert databases from one format or type to another.

Source: http://www.webopedia.com/TERM/E/ETL.html

Covey Principle: Physical representation of the Data lake

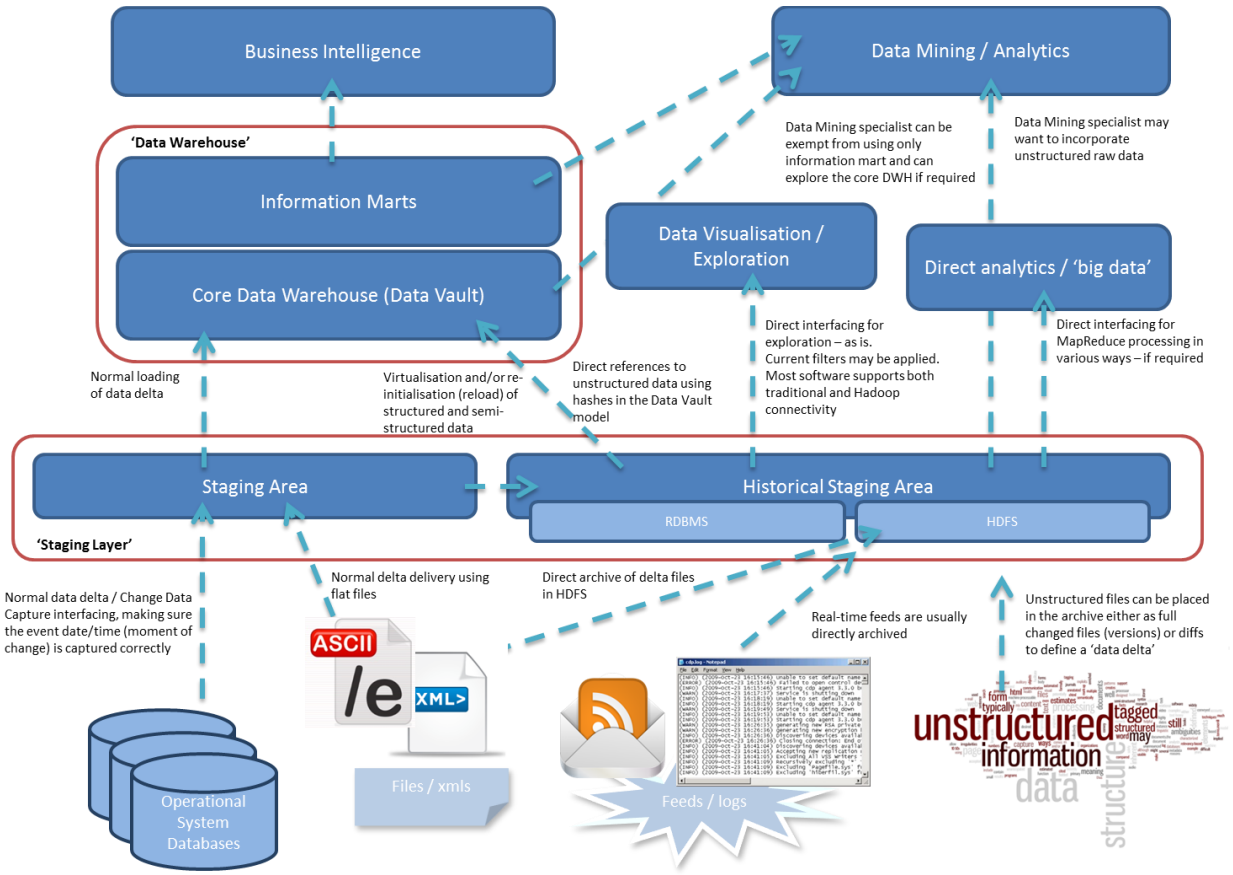
Every company defines the data lake differently.

A data lake is a storage repository that holds a vast amount of raw data in its native format until it is needed. While a hierarchical data warehouse stores data in files or folders, a data lake uses a flat architecture to store data.

Source: http://searchaws.techtarget.com/definition/data-lake

A representation of the Data lake

Source: http://roelantvos.com/blog/?page_id=1375



Components of the Data Lake:

- Data movement
- Data access
- Processing engines
- Dataflow tools
- Scheduling & workflow management
- Metadata
- Data curation
- Platform services

We also discussed the importance of the V's of "Big Data":

http://www.optimusinfo.com/understanding-the-7-vs-of-big-data/

Tips for Project:

Think about the following:

- Type of data you are using
- Where the data comes from
- Problem you will solve
- Where in the company you will solve it

Characteristics for a Modern Database Architecture:

Work with business users to identify the types of data that are most valuable. Info needs high business impact

Make data governance a first priority- identifying, ingesting, building models for data needs to assure quality, and relevance for the business. Responsibility for the data

Build systems to change not to last. The architecture needs to accommodate solutions

Develop a real-time foundation

Build security into the foundation, data security

Developer a master data management strategy

Position data as a service

Offer self-service environments. Don't want to have to wait for IT or data management departments to deliver the information

Sources:

Database Trend & Applications, Jan 8, 2015

http://www.dbta.com/Editorial/Think-About-It/8-Steps-to-Building-a-Modern-Data-Architecture-101417.aspx