Different Shapes of Data

COURSE DESCRIPTION

For several decades the world data was characterized by one strong constant — we store data in relational databases. Since 2010 this landscape changed radically. One can argue about increased volumes or the users' appetite for higher variety of data, but the best way to catch the change is to see what business problems are solved with the help of this new landscape.

To understand the magnitude of this change, here is one number: the market once dominated by a few dozen of data systems, now has more than 300 ranked members, 60% of these represents new NoSQL solutions.

There is a lot value in such a variety, but also challenge: the world has shifted from general solution to specialized solution, from closed commercial systems to open source systems, therefore choosing the right solution to the problem became a key element of the business analytics.

This course meant to be a practical course presenting several business scenarios and the appropriate data solutions to support these scenarios. From the simplest relational cubes we will get to the "different shapes of data", the immense variety of new technologies and data concepts all meant to support a new world of information. We will not focus much on individual tools but complete data processes, from the acquisition of the data until the birth of the business value.

COURSE OBJECTIVES

The aims of this course:

- The get the students familiar with different aspects of the data world, present an overview of the major data technologies, concepts and study them in the context of business problems
- To stress the importance of business value in the world of data
- To connect the dots: during the Business Analytics program the students have the opportunity to learn several techniques including Big Data, Machine Learning or Exploratory Data Analysis. The aim this course is to link these techniques in complex data processes (It is however not a prerequisite to know these techniques in advance to complete the course)

The learning outcome of the course:

At the end of the course, students will be able to choose a proper data strategy for a business scenario by knowing the options and limitations.

HOW THE CLASS SESSIONS WILL BE CONDUCTED

Each lecture will have theoretical introduction followed by a use case presented sometimes by guests. This is not a lab course however, code samples will be presented in some cases to show solutions in practice, these samples will be used later by student to solve the home assignment.

SESSIONS

1. THE VALUE IN DATA & BUSINESS INTELLIGENCE
   Business organizations and data. ERP. Business Intelligence. Business Analytics. Data OLTP. Data Warehouses. ELT. OLAP.

2. THE EMERGE OF THE NEW QUESTIONS AND CHALLENGES
   Old and new business questions. Challenges with RBDMS. Scaling. CAP Theorem. ACID. High Availability. Fault tolerant systems. Replication. The new tools. Cloud. DBaaS.

3. INTRODUCING THE NEW PLATFORMS OF DATA ANALYTICS
   New trends. Analytical vs operational data problems. NoSQL in brief. Db-engines.com. Key Value Stores (Redis, Dynamo). Demo. Use case review. Data pipelines. Lambda architecture.

4. ADVANCED SEARCH
   Wide column stores (Cassandra). Demo. Use case review. Search engines. (Solr). Full text search. Geo search. Faceted search. Demo. Use case review.

5. SPECIAL TOOLS FOR SPECIAL PROBLEMS
   Time series analytics. Time Series DBMS (InfluxDB). Demo. Use case review. Graph theory/Network science. Graph DBMS (Neo4J). Demo. Use case review. Geographical Information Systems.

6. THE BIG PLAYER & POLYGLOT PERSISTENCY
   Document Stores (MongoDB). Relation to other stores. Demo. Use case review. Polyglot persistency. Best Practices.

ASSESSMENT

Quizzes with Socrative (10%)
Participation during class sessions (10%)
Home exercise (30%)
Use case presentation (50%)

USE CASE PRESENTATION

Find a real world business use case with a complex data process implementation. Create a detailed presentation about it, using the concepts learned in the classroom. What is the business benefit? What about the ROI? Would you recommend any alternative solution to this problem?

The presentation: max 10 minutes + 2 minutes for questions