About the course

Overview

This course is about big data management, focussing on how to store big data, and how to analyse it. The aim is to prepare you for discussions that you might have in your organisation, about storing and analysing your big data, and to help you make more informed decisions.

There are many concepts and technologies involved in the management of big data. You will learn about these at a high level, getting a clear view of the big picture. But you will also get some hands-on experience working with the technologies themselves. This is not to help you develop any deep expertise in those technologies, but rather to help you better understand the underlying concepts and issues - it is these underlying concepts and issues that we are primarily concerned with, and often the best way to properly grasp them is to get some real, hands-on experience. It is also a great way to get a proper feel for what these technologies can and cannot do, and why you might choose one over another.

Weekly topics

In **Week 1** we lay the foundations. We discuss the different types of data format and their degree of structure. We look at the units involved in measuring data volume, and what they mean. We consider when data counts as big data, in terms of volume, variety, and velocity. We discuss the value of big data to an organisation. We look at the various ways of storing data (relational vs non-relational databases, single machine vs network of machines) and when to choose one over another. We consider some of the challenges involved in storing and manipulating big data.

We also focus on relational databases. We look at what a relational database is. We draw a distinction between two kinds of relational databases - operational databases, and data warehouses (the latter are an important part of the management of big data) - and discuss when you might use one rather than the other. We discuss how to design a relational database, and how the best way to do so differs between operational databases and data warehouses.

In **Week 2** you will learn how to extract data from a relational database, using Structured Query Language (SQL). SQL is the primary means of managing data in a relational database. In particular, it is the primary means of extracting data to answer organisational questions. It is thus important for you to have a good sense of what SQL is and how it works. Moreover, SQL is becoming more commonly used to manage data in non-relational databases as well, making it more important for you to have an understanding of SQL.

In **Week 3** we turn to the problem of storing and working with massive volumes of data. One of the main techniques for storing such big data is to use a file management system called Hadoop, which manages the distribution and replication of data across a network of computers (potentially thousands of them). You will learn about the Hadoop file management system, and get some hands-on experience using it. Along with Hadoop comes a fundamentally important technique for analysing distributed data, called MapReduce. You will learn about and use MapReduce (which will get you using your Python skills). You'll also learn about the numerous pieces of software that have been developed to help people use Hadoop and MapReduce, forming what's called "The Hadoop Ecosystem".

In **Week 4** you'll learn about and use MRJob, a Python Library for MapReduce programming, that has been developed to simplify the process of analysing big data on Hadoop. MRJob is the easiest route to writing Python programs that run on Hadoop. If you use MRJob, you'll be able to test your code locally without installing Hadoop or run it on a cluster of your choice.

In **Week 5** you will learn about and use Spark, another piece of software that has been developed to simplify, and also speed up, the analysis of big data on Hadoop. Because of its speed and ease of use, Spark is becoming one of the most important tools in the management of big data. We will learn two types of Spark APIs this week: RDD and DataFrame.

In **Week 6** it is where your SQL skills will come back in, and we will learn two tools for big data

management with SQL-like interfaces: Hive and Spark SQL. Hive allows you to use SQL to query massive volumes of data, even though it is not stored in a relational database. Similarly, Spark SQL allows you to execute SQL queries in Spark. It means that we even do not need to learn to program over MapReduce and Spark, but we can still enjoy their power of them in big data management using SQL.

Approach to learning

Each week there are some **theory slides** for you to work through, and in most weeks there are also some **coding examples**. Together these cover the concepts and skills for the week and should be your primary source of learning.

In addition, there are a variety of **videos** for you to watch. These are all freely available on YouTube, and we've added them here because you'll probably find them helpful supplements to the theory slides and coding examples. As you have probably discovered, there are many excellent videos on YouTube. But, as you've probably also discovered, it can be difficult to find them among myriad other ones. We've done some of that work for you - these are videos whose content we endorse and that we think you might find helpful. Some of them overlap, but their styles are diverse, so by watching them you'll experience a good range of explanations.

You will also find some **further resources** slides, which contain suggestions about which resources to use if you want to learn more.

Each week there is an **assessment task**. This aims to test your understanding of the main concepts and/or skills for the week.

Useful Resources

The course has used some materials from the following textbooks:

1. Fundamentals of database systems

This book covers most topics about RDBMS. This course only covers the topics of the first few chapters in this book. If you would like to learn more about RDBMS and SQL, you can read the latter chapters on your own.

2. Hadoop The Definitive Guide

This books is kind of an encyclopaedia of the Hadoop Ecosystem, and it covers many Hadoop projects, including HDFS, MapReduce, Spark, Hive, HBase, etc. It is not recommended to read the whole book for learning this course. You can refer to this book when you have some questions.

3. Learning Spark, 1st edition

This book introduces how to use Spark RDDs to solve big data problems.

4. Learning Spark, 2nd edition

This book introduces how to use Spark's new structured APIs (DataSet and DataFrame) to solve big data problems.

Due to the time limit, this course can only cover a few big data management techniques. If you would like to learn more on your own, the above four books would be very helpful.