Yandex Coursera Study Notes

- Coursera course at: https://www.coursera.org/learn/big-data-essentials

Historical data
Master of information is master of situation.

Reduction of amount of rejects in production

Gathering
Keeping
Analyzing

Information extraction
What are big data essential skills?

Ability to work with Hadoop HDFS
10,000 nodes – 10 nodes will break everyday

large files are stored by chunks.
MapReduce
How to transform a computation into chunks of MapReduce tasks

Different frameworks that use MapReduce under the hood.

Apache Spark
In-Memory framework

Design distributed file system
RDD
Apache Spark

Data Engineers

Big Data: task is challenging
6 weeks
2nd week of course: MapReduce
1 week for programming assignments

Apache Spark: modern computation framework
Practical assignments

Last week: case studies

How to work with sample data

Iterative algorithms
Finding shortest path in social graph

Chief Data Scientist at Yandex.

Machine learning specialization

Moscow State University

Junior researcher
Yandex
Amazon AWS

Big Data & Machine Learning

Essential to get knowledge

Ivan: Big Data Essentials

Topic on Spark
Major in math

5 years: Yandex
petabytes of data

scalability issues

operational mechanics
Yandex School of Analysis

File Systems:

How to reach it fast
Store data into the named files
Hierarchy of files
Structure and logic rules used to manage files

File Manager
GNOME: Nautilus
Windows Explorer
Finder

Bash:
Command interpreter
Programming language
Command processor that typically runs in a text window, where the user types command that cause
Actions
Can also read and execute commands from a file, called script

du command

df command