

The Data Engineering Cookbook

How to master the plumbing of data science

Andreas Kretz

December 2, 2018

v0.1

Contents

1	Introduction	6
2	What is Data Science?	7
2.1	Data Scientist	7
2.2	Data Engineer	7
2.3	Data Analyst	7
3	The Basic Skills	7
4	Learn to Write Code	7
4.1	Coding Basics	8
4.2	Learn To Use GitHub	8
4.3	Agile Development	9
4.3.1	Scrum	9
4.3.2	OKR	9
5	Computer Science Basics	9
5.1	Learn how a Computer Works	9
5.1.1	CPU,RAM,GPU,HDD	9
5.1.2	Differences between PCs and Servers	9
5.2	Computer Networking	9
5.2.1	ISO/OSI Model	9
5.2.2	IP Subnetting	9
5.2.3	Switch, Level 3 Switch	9
5.2.4	Router	9
5.2.5	Firewalls	9
5.3	Security and Privacy	10
5.3.1	SSL Public & Private Key Certificates	10
5.3.2	What is a certificate authority	10
5.3.3	JAVa Web Tokens	10
5.3.4	GDPR regulations	10
5.3.5	Privacy by design	10
5.4	Linux	10
5.4.1	OS Basics	10
5.4.2	Shell scripting	10
5.4.3	Cron jobs	10
5.4.4	Packet management	10

6	Data Science Platform	10
6.1	Security Zone Design	10
6.1.1	How to secure a multi layered application	10
6.1.2	Cluster security with Kerberos	10
6.2	Lambda Architecture	11
6.2.1	Stream and Batch processing	11
6.2.2	My Big Data Platform Blueprint	11
6.2.3	Three methods of streaming	13
6.3	Big Data	15
6.3.1	What is big data and where is the difference to data science and data analytics?	15
6.3.2	The 4Vs of Big data:	15
6.3.3	Why Big Data?	16
6.3.4	What are the tools associated?	20
6.4	What is the difference between a Data Warehouse and a Data Lake . . .	20
6.5	Hadoop Platfroms	20
6.5.1	What makes Hadoop so popular?	21
6.5.2	How does a Hadoop System architecture look like	24
6.5.3	What tools are usually in a with Hadoop Cluster	24
6.5.4	How to select Hadoop Cluster Hardware	24
6.6	Is ETL still relevant for Analytics?	24
6.7	The Cloud	25
6.7.1	AWS,Azure, IBM, Google Cloud basics	25
6.7.2	cloud vs on premise	25
6.7.3	up & downsides	25
6.7.4	Security	25
6.8	Docker	25
6.8.1	What is docker and what do you use it for	25
6.8.2	How to create, start,stop a Container	27
6.8.3	Docker micro services?	27
6.8.4	Kubernetes	27
6.8.5	Why and how to do Docker container orchestration	27
7	How to Ingest Data	27
7.1	Application programming interfaces	27
7.1.1	REST APIs	27
7.1.2	HTTP Post/Get	27
7.1.3	API Design	27
7.1.4	Implementation	27

7.1.5	OAuth security	27
7.2	JSON	27
7.2.1	Super important for REST APIs	27
7.2.2	How is it used for logging and processing logs	27
8	Distributed Processing	27
8.1	MapReduce	28
8.1.1	Why was MapReduce Invented	28
8.1.2	How does that work	28
8.1.3	What is the limitation of MapReduce?	28
8.2	Apache Spark	28
8.2.1	Spark Basics	29
8.2.2	What is the difference to MapReduce?	29
8.2.3	How to do stream processing	29
8.2.4	How to do batch processing	29
8.2.5	How does Spark use data from Hadoop	29
8.2.6	What is a RDD and what is a DataFrame?	29
8.2.7	Spark coding with Scala	29
8.2.8	Spark coding with Python	29
8.2.9	How and why to use SparkSQL?	29
8.2.10	Machine Learning on Spark? (Tensor Flow)	29
8.3	Message queues with Apache Kafka	29
8.3.1	Why a message queue tool?	29
8.3.2	Kakfa architecture	29
8.3.3	What are topics	29
8.3.4	What does Zookeeper have to do with Kafka	29
8.3.5	How to produce and consume messages	29
8.4	Machine Learning	29
8.4.1	Training and Applying models	30
8.4.2	What is deep learning	30
8.4.3	How to do Machine Learning in production	30
8.4.4	Why machine learning in production is harder then you think	30
8.4.5	How to convince people machine learning works	31
9	How to Store Data	33
9.1	Data Modeling	34
9.1.1	How to find out how you need to store data for the business case	34
9.1.2	How to decide what kind of storage you need to use	34

9.2	SQL	34
9.2.1	Database Design	34
9.2.2	SQL Queries	34
9.2.3	Stored Procedures	34
9.2.4	ODBC/JDBC Server Connections	34
9.3	NoSQL	34
9.3.1	KeyValue Stores (HBase)	34
9.3.2	Document Stores (HDFS, MongoDB)	34
9.3.3	Hive Warehouse	36
9.3.4	Impala	36
9.3.5	Time Series Databases (?)	36
9.3.6	MPP Databases (Greenplum)	37
10	How to Visualize Data	37
10.1	Mobile Apps	37
10.1.1	Android & IOS basics	37
10.1.2	How to design APIs for mobile apps	37
10.2	How to use Webservers to display content	37
10.2.1	Tomcat	38
10.2.2	Jetty	38
10.2.3	NodeRED	38
10.2.4	React	38
10.3	Business Intelligence Tools	38
10.3.1	Tableau	38
10.3.2	PowerBI	38
10.3.3	Quliksense	38
10.4	Identity & Device Management	38
10.4.1	What is a digital twin?	38
10.4.2	Active Directory	38
11	Case Studies	38
11.1	Data Science @Airbnb	38
11.2	Data Science @Netflix	38
11.3	Data Science @Uber	38
11.4	Data Scieene @Zalando	38
12	DEV/OPS	38
12.1	Hadoop	38
12.1.1	Hadoop Cluster setup and management with Cloudera Manager (for example)	39

12.1.2 Spark code from coding to production	39
12.1.3 How to monitor and manage data processing pipelines	39
12.1.4 Oozie	39
12.1.5 Airflow Application management	39
12.1.6 Creating Statistics with Spark and Kafka	39

1 Introduction

What do you actually need to learn to become an awesome data engineer? Look no further, you find it here.

How to use this document: This is not a training! It's a collection of skills, that I value highly in my daily work as a data engineer. It's intended to be a starting point for you to find the topics to look into.

This project is a work in progress! Over the next weeks I am going to share with you my thoughts on why each topic is important. I also try to include links to useful resources.

How to find out what is new? You will always find the newest version on my Patreon **<https://www.patreon.com/plumbersofds>**

Help make this collection awesome! Join the discussion on Patreon or write me an email to andreaskayy@gmail.com. Tell me your thoughts, what you value, you think should be included, or where I am wrong.

– Andreas