

Contents

1	How To Use This Cookbook	6
2	Data Engineer vs Data Scientists	7
2.1	Data Scientist	7
2.2	Data Engineer	8
2.3	Who Companies Need	9
3	Data Engineering Example	9
4	Building A Data Platform	9
4.1	Lambda Architecture	9
4.1.1	Batch Processing	9
4.1.2	Stream Processing	10
4.2	My Big Data Platform Blueprint	11
4.2.1	Ingest	12
4.2.2	Analyse / Process	13
4.2.3	Store	13
4.2.4	Display	14
4.3	Lambda Architecture Alternative	15
4.3.1	Kappa Architecture	15
4.3.2	Kappa Architecture with Kudu	15
4.4	Thoughts On Choosing The Target Environment	15
4.4.1	Cloud vs On-Premise	15
4.4.2	Cloud Native or Independent Vendors	15
4.5	Thoughts On Choosing A Development Environment	15
4.5.1	Cloud As Dev Environment	15
4.5.2	Local Dev Environment	15
5	Data Architecture	15
5.1	Source Data	15
5.2	Analytics Requirements For Streaming	15
5.3	Analytics Requirements For Batch Processing	15
5.4	Data Visualization	15
6	Milestone 1 — Tool Decisions	15
7	Basic Data Engineering Skills	15
7.1	Coding Basics	16
7.2	Learn To Use GitHub — available	16

7.3	Agile Development – available	17
7.3.1	Agile rules I learned over the years – available	18
7.3.2	Scrum	20
7.3.3	OKR	20
7.4	Learn how a Computer Works	20
7.4.1	CPU,RAM,GPU,HDD	20
7.4.2	Differences between PCs and Servers	20
7.5	Computer Networking	20
7.5.1	ISO/OSI Model	20
7.5.2	IP Subnetting	20
7.5.3	Switch, Level 3 Switch	20
7.5.4	Router	20
7.5.5	Firewalls	20
7.6	Security and Privacy	21
7.6.1	SSL Public & Private Key Certificates	21
7.6.2	What is a certificate authority	21
7.6.3	JAVA Web Tokens	21
7.6.4	GDPR regulations	21
7.6.5	Privacy by design	21
7.7	Linux	21
7.7.1	OS Basics	21
7.7.2	Shell scripting	21
7.7.3	Cron jobs	21
7.7.4	Packet management	21
7.8	The Cloud	21
7.8.1	AWS,Azure, IBM, Google Cloud basics	21
7.8.2	cloud vs on premise	21
7.8.3	up & downsides	21
7.8.4	Security	21
7.9	Security Zone Design	22
7.9.1	How to secure a multi layered application	22
7.9.2	Cluster security with Kerberos	22
7.9.3	Kerberos Tickets	22
7.10	Three methods of streaming — available	22
7.11	Big Data	24
7.11.1	What is big data and where is the difference to data science and data analytics?	24
7.11.2	The 4Vs of Big Data — available	24
7.11.3	Why Big Data? — available	25

7.11.4	What are the tools associated?	29
7.12	What is the difference between a Data Warehouse and a Data Lake . . .	29
7.13	Hadoop Platforms — available	29
7.13.1	What makes Hadoop so popular? — available	29
7.13.2	How does a Hadoop System architecture look like	33
7.13.3	What tools are usually in a with Hadoop Cluster	33
7.13.4	How to select Hadoop Cluster Hardware	33
7.14	Is ETL still relevant for Analytics?	33
7.15	Docker	33
7.15.1	What is docker and what do you use it for — available	33
7.15.2	How to create, start,stop a Container	35
7.15.3	Docker micro services?	35
7.15.4	Kubernetes	35
7.15.5	Why and how to do Docker container orchestration	35
7.16	REST APIs	35
7.16.1	HTTP Post/Get	35
7.16.2	API Design	35
7.16.3	Implementation	35
7.16.4	OAuth security	35
7.17	Data Modeling	35
7.17.1	How to find out how you need to store data for the business case .	35
7.17.2	How to decide what kind of storage you need to use	35
7.18	SQL Databases	35
7.19	NoSQL Stores	36
7.19.1	KeyValue Stores (HBase)	36
7.19.2	Document Store HDFS — available	36
7.19.3	Document Store MongoDB	38
7.19.4	Hive Warehouse	38
7.19.5	Impala	38
7.19.6	Time Series Databases	38
7.19.7	MPP Databases (Greenplum)	38
7.20	MapReduce – available	38
7.21	Apache Spark	42
7.21.1	Spark Basics	42
7.22	Apache Kafka	47
7.23	Machine Learning	47
7.24	Data Visualization	51
7.24.1	Mobile Apps	51
7.24.2	How to use Webservers to display content	51

7.24.3	Business Intelligence Tools	51
7.24.4	Identity & Device Management	52
8	Case Studies	52
8.1	7 Steps to Successfull Data Science Project	52
8.2	Data Science @Airbnb	52
8.3	Data Sciecne @Baidu	52
8.4	Data Sciecne @Blackrock	52
8.5	Data Sciecne @BMW	52
8.6	Data Sciecne @Booking.com	53
8.7	Data Sciecne @CERN	53
8.8	Data Sciecne @Disney	53
8.9	Data Sciecne @Drivetribe	54
8.10	Data Sciecne @Dropbox	54
8.11	Data Sciecne @Ebay	54
8.12	Data Sciecne @Expedia	54
8.13	Data Sciecne @Facebook	54
8.14	Data Sciecne @@Grammarly	54
8.15	Data Sciecne @ING Fraud	54
8.16	Data Sciecne @Instagram	54
8.17	Data Sciecne @LinkedIn	55
8.18	Data Sciecne @Lyft	55
8.19	Data Sciecne @NASA	55
8.20	Data Science @Netflix – available	55
8.21	Data Sciecne @OTTO	59
8.22	Data Sciecne @Paypal	59
8.23	Data Sciecne @Pinterest	59
8.24	Data Sciecne @Salesforce	59
8.25	Data Sciecne @Slack	60
8.26	Data Sciecne @Spotify	60
8.27	Data Sciecne @Symantec	60
8.28	Data Science @Tinder	60
8.29	Data Science @Twitter	60
8.30	Data Science @Uber	60
8.31	Data Science @Upwork	60
8.32	Data Sciecne @Woot	61
8.33	Data Sciecne @Zalando	61