# The Data Engineering Cookbook

### Mastering the Plumbing of Data Science

IZZAT FAIZ JOHAN

Aug 2019

# Contents

I Introduction [10](#_bookmark0)

1. How To Use This Cookbook [11](#_bookmark1)
2. Data Engineer vs Data Scientists [12](#_bookmark2)
   1. Data Scientist [12](#_bookmark4)
   2. Data Engineer [13](#_bookmark6)
   3. Who Companies Need [14](#_bookmark8)
3. Basic Data Engineering Skills [16](#_bookmark9)
4. The Cloud [30](#_bookmark44)
   1. IaaS vs PaaS vs SaaS [30](#_bookmark45)
   2. AWS,Azure, IBM, Google Cloud basics [30](#_bookmark47)
   3. Cloud vs On-Premises [30](#_bookmark48)
   4. Security [31](#_bookmark50)
   5. Hybrid Clouds [31](#_bookmark51)
5. Security Zone Design [32](#_bookmark52)
   1. How to secure a multi layered application [32](#_bookmark53)
   2. Cluster security with Kerberos [32](#_bookmark54)
   3. Kerberos Tickets [32](#_bookmark55)
6. Big Data [33](#_bookmark56)
   1. What is big data and where is the difference to data science and data analytics? [33](#_bookmark57)
   2. The 4Vs of Big Data — available [33](#_bookmark58)
   3. Why Big Data? — available [34](#_bookmark59)
      1. Planning is Everything [35](#_bookmark60)
      2. The Problem with ETL [35](#_bookmark61)
      3. Scaling Up [36](#_bookmark63)
      4. Scaling Out [37](#_bookmark65)
      5. Please don go Big Data [38](#_bookmark67)
7. My Big Data Platform Blueprint [39](#_bookmark68)
   1. Ingest [40](#_bookmark70)
   2. Analyse / Process [40](#_bookmark71)
   3. Store [41](#_bookmark72)
   4. Display [42](#_bookmark73)
8. Lambda Architecture [43](#_bookmark74)
   1. Batch Processing [43](#_bookmark76)
   2. Stream Processing [44](#_bookmark78)
   3. Should you do stream or batch processing? [45](#_bookmark80)
   4. Lambda Architecture Alternative [45](#_bookmark81)
      1. Kappa Architecture [45](#_bookmark82)
      2. Kappa Architecture with Kudu [45](#_bookmark83)
   5. Why a Good Data Platform Is Important [45](#_bookmark84)
9. Data Warehouse vs Data Lake [46](#_bookmark86)
10. Hadoop Platforms — available [47](#_bookmark88)
    1. What is Hadoop [47](#_bookmark89)
    2. What makes Hadoop so popular? — available [48](#_bookmark91)
    3. Hadoop Ecosystem Components [49](#_bookmark93)
    4. Hadoop Is Everywhere? [50](#_bookmark96)
    5. Should you learn Hadoop? [51](#_bookmark97)

How does a Hadoop System architecture look like [51](#_bookmark98)

What tools are usually in a with Hadoop Cluster [51](#_bookmark99)

* 1. How to select Hadoop Cluster Hardware [51](#_bookmark100)

1. Docker [52](#_bookmark101)
   1. What is docker and what do you use it for — available [52](#_bookmark102)
      1. Don’t Mess Up Your System [52](#_bookmark103)
      2. Preconfigured Images [52](#_bookmark104)
      3. Take It With You [53](#_bookmark105)
   2. Kubernetes Container Deployment [53](#_bookmark106)
   3. How to create, start,stop a Container [54](#_bookmark107)
   4. Docker micro services? [54](#_bookmark108)
   5. Kubernetes [54](#_bookmark109)
   6. Why and how to do Docker container orchestration [54](#_bookmark110)
   7. Useful Docker Commands [54](#_bookmark111)
2. REST APIs [56](#_bookmark112)
   1. API Design [56](#_bookmark114)
   2. Implementation Frameworks [56](#_bookmark116)
   3. OAuth security [57](#_bookmark117)
3. Databases [58](#_bookmark118)
   1. SQL Databases [58](#_bookmark119)
      1. PostgreSQL DB [58](#_bookmark120)
      2. Database Design [58](#_bookmark121)
      3. SQL Queries [58](#_bookmark122)
      4. Stored Procedures [58](#_bookmark123)
      5. ODBC/JDBC Server Connections [58](#_bookmark124)
   2. NoSQL Stores [58](#_bookmark125)
      1. KeyValue Stores (HBase) [58](#_bookmark126)
      2. Document Store HDFS — available [58](#_bookmark127)
      3. Document Store MongoDB [61](#_bookmark131)
      4. Elasticsearch Search Engine and Document Store [62](#_bookmark133)
      5. Hive Warehouse [63](#_bookmark135)
      6. Impala [63](#_bookmark136)
      7. Kudu [63](#_bookmark137)
      8. Apache Druid [63](#_bookmark138)
      9. InfluxDB Time Series Database [63](#_bookmark140)
      10. MPP Databases (Greenplum) [64](#_bookmark141)
4. Data Processing and Analytics - Frameworks [65](#_bookmark142)
   1. Is ETL still relevant for Analytics? [65](#_bookmark143)
   2. Stream Processing [65](#_bookmark145)
      1. Three methods of streaming — available [65](#_bookmark146)
      2. At Least Once [66](#_bookmark147)
      3. At Most Once [66](#_bookmark148)
      4. Exactly Once [66](#_bookmark149)
      5. Check The Tools! [67](#_bookmark150)
   3. MapReduce [67](#_bookmark151)
      1. How does MapReduce work – available [68](#_bookmark153)
      2. Example [69](#_bookmark154)
      3. What is the limitation of MapReduce? – available [71](#_bookmark156)
   4. Apache Spark [71](#_bookmark158)
      1. What is the difference to MapReduce? – available [71](#_bookmark159)
      2. How does Spark fit to Hadoop? – available [72](#_bookmark160)
      3. Where’s the difference? [72](#_bookmark162)
      4. Spark and Hadoop is a perfect fit [73](#_bookmark163)
      5. Spark on YARN: [73](#_bookmark164)
      6. My simple rule of thumb: [74](#_bookmark165)
      7. Available Languages – available [74](#_bookmark166)
      8. How to do stream processing [74](#_bookmark167)
      9. How to do batch processing [74](#_bookmark168)
      10. How does Spark use data from Hadoop – available [74](#_bookmark169)
      11. What is a RDD and what is a DataFrame? [75](#_bookmark171)
      12. Spark coding with Scala [75](#_bookmark172)
      13. Spark coding with Python [75](#_bookmark173)
      14. How and why to use SparkSQL? [75](#_bookmark174)
      15. Machine Learning on Spark? (Tensor Flow) [75](#_bookmark175)
      16. MLlib: [75](#_bookmark176)
      17. Spark Setup – available [76](#_bookmark177)
      18. Spark Resource Management – available [76](#_bookmark178)
   5. Apache Nifi [76](#_bookmark179)
   6. StreamSets [76](#_bookmark180)
5. Apache Kafka [78](#_bookmark182)
   1. Why a message queue tool? [78](#_bookmark183)
   2. Kakfa architecture [78](#_bookmark184)
   3. What are topics [78](#_bookmark185)
   4. What does Zookeeper have to do with Kafka [78](#_bookmark186)
   5. How to produce and consume messages [78](#_bookmark187)
   6. KAFKA Commands [78](#_bookmark188)
6. Machine Learning [80](#_bookmark189)
   1. Training and Applying models [80](#_bookmark191)
   2. What is deep learning [80](#_bookmark192)
   3. How to do Machine Learning in production — available [80](#_bookmark193)
   4. Why machine learning in production is harder then you think – available [81](#_bookmark194)
   5. Models Do Not Work Forever [81](#_bookmark195)
   6. Where The Platforms That Support This? [81](#_bookmark196)
   7. Training Parameter Management [82](#_bookmark197)
   8. What’s Your Solution? [82](#_bookmark198)
   9. How to convince people machine learning works — available [82](#_bookmark199)
   10. No Rules, No Physical Models [83](#_bookmark200)
   11. You Have The Data. USE IT! [83](#_bookmark201)
   12. Data is Stronger Than Opinions [84](#_bookmark202)
   13. AWS Sagemaker [84](#_bookmark203)
7. Data Visualization [85](#_bookmark204)
   1. Android & IOS [85](#_bookmark205)
   2. How to design APIs for mobile apps [85](#_bookmark206)
   3. How to use Webservers to display content [85](#_bookmark207)
      1. Tomcat [86](#_bookmark208)
      2. Jetty [86](#_bookmark209)
      3. NodeRED [86](#_bookmark210)
      4. React [86](#_bookmark211)
   4. Business Intelligence Tools [86](#_bookmark212)
      1. Tableau [86](#_bookmark213)
      2. PowerBI [86](#_bookmark214)
      3. Quliksense [86](#_bookmark215)
   5. Identity & Device Management [86](#_bookmark216)
      1. What is a digital twin? [86](#_bookmark217)
      2. Active Directory [86](#_bookmark218)
8. Data Engineering Course: Building A Data Platform [87](#_bookmark219)
9. What We Want To Do [88](#_bookmark220)
10. Thoughts On Choosing A Development Environment [89](#_bookmark221)
11. A Look Into the Twitter API [90](#_bookmark223)
12. Ingesting Tweets with Apache Nifi [91](#_bookmark225)
13. Writing from Nifi to Apache Kafka [92](#_bookmark228)
14. Apache Zeppelin [93](#_bookmark231)
    1. Install and Ingest Kafka Topic [93](#_bookmark232)
    2. Processing Messages with Spark & SparkSQL [93](#_bookmark233)
    3. Visualizing Data [93](#_bookmark234)
15. Switch Processing from Zeppelin to Spark [94](#_bookmark235)
    1. Install Spark [94](#_bookmark236)
    2. Ingest Messages from Kafka [94](#_bookmark237)
    3. Writing from Spark to Kafka [94](#_bookmark238)
    4. Move Zeppelin Code to Spark [94](#_bookmark239)
16. Case Studies [95](#_bookmark240)
17. How I do Case Studies [96](#_bookmark241)
    1. Data Science @Airbnb [96](#_bookmark242)
       1. Data Science @Amazon [96](#_bookmark244)
    2. Data Science @Baidu [96](#_bookmark245)
    3. Data Science @Blackrock [97](#_bookmark246)
    4. Data Science @BMW [97](#_bookmark247)
    5. Data Science @Booking.com [97](#_bookmark248)
    6. Data Science @CERN [97](#_bookmark250)
    7. Data Science @Disney [98](#_bookmark252)
    8. Data Science @Drivetribe [98](#_bookmark253)
    9. Data Science @Dropbox [98](#_bookmark254)
    10. Data Science @Ebay [99](#_bookmark255)
    11. Data Science @Expedia [99](#_bookmark256)
    12. Data Science @Facebook [99](#_bookmark257)
    13. Data Science @Google [99](#_bookmark258)
    14. Data Science @@Grammarly [99](#_bookmark259)
    15. Data Science @ING Fraud [99](#_bookmark260)
    16. Data Science @Instagram [99](#_bookmark261)
    17. Data Science @LinkedIn [100](#_bookmark262)
    18. Data Science @Lyft [100](#_bookmark264)
    19. Data Science @NASA [100](#_bookmark265)
    20. Data Science @Netflix – available [101](#_bookmark267)
    21. Data Science @OLX [105](#_bookmark272)
    22. Data Science @OTTO [105](#_bookmark274)
    23. Data Science @Paypal [105](#_bookmark275)
    24. Data Science @Pinterest [105](#_bookmark276)
    25. Data Science @Salesforce [106](#_bookmark278)
    26. Data Science @Siemens Mindsphere [106](#_bookmark279)
    27. Data Science @Slack [106](#_bookmark281)
    28. Data Science @Spotify [107](#_bookmark282)
    29. Data Science @Symantec [107](#_bookmark284)
    30. Data Science @Tinder [107](#_bookmark285)
    31. Data Science @Twitter [107](#_bookmark286)
    32. Data Science @Uber [108](#_bookmark288)
    33. Data Science @Upwork [109](#_bookmark289)
    34. Data Science @Woot [109](#_bookmark290)
    35. Data Science @Zalando [109](#_bookmark291)
18. 1001 Data Engineering Interview Questions [111](#_bookmark293)
19. Live Streams [113](#_bookmark294)
20. All Interview Questions [114](#_bookmark296)

**Part I Introduction**

# How To Use This Cookbook

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

If you are looking for AI algorithms and such data scientist things, this book is not for you.

#### How to use this document:

First of all, this is not a training! This cookbook is 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 and become an awesome data engineer.

You are going to find ***Five Types of Content*** in this book: Articles I wrote, links to my podcast episodes (video & audio), more then 200 links to helpful websites I like, data engineering interview questions and case studies.

#### This book is a work in progress!

As you can see, this book is not finished. I’m constantly adding new stuff and doing videos for the topics. But obviously, because I do this as a hobby my time is limited. You can help making this book even better.

#### Help make this book awesome!

If you have some cool links or topics for the cookbook, please become a contributor on GitHub: <https://github.com/andkret/Cookbook>. Pull the repo, add them and create a pull request. Or join the discussion by opening Issues. You can also write me an email any time to [plumbersofdatascience@gmail.com.](mailto:plumbersofdatascience@gmail.com) Tell me your thoughts, what you value, you think should be included, or correct where I am wrong.

#### This Cookbook is and will always be free!

I don’t want to sell you this book, but please support what you like and join my Patreon: <https://www.patreon.com/plumbersofds>

Check out this podcast episode where I talk in detail why I decided to share all this information for free: [#079 Trying to stay true to myself and making the cookbook public](https://youtu.be/k1bS5aSPos8) [on GitHub](https://youtu.be/k1bS5aSPos8)

# Data Engineer vs Data Scientists

**Podcast Episode:** #050 Data Engineer Scientist or Analyst Which One Is For You?

In this podcast we talk about the differences between data scientists, analysts and engineers. Which are the three main data science jobs. All three super important. This makes it easy to decide

YouTube [Click here to watch](https://youtu.be/64TYZETOEdQ) Audio [Click here to listen](https://anchor.fm/andreaskayy/episodes/050-Data-Engineer-Scientist-or-Analyst-Which-One-Is-For-You-e45ibl)

Table 2.1: Podcast: 050 Data Engineer Scientist or Analyst Which One Is For You?

## Data Scientist

Data scientists aren’t like every other scientist.

Data scientists do not wear white coats or work in high tech labs full of science fiction movie equipment. They work in offices just like you and me.

What differs them from most of us is that they are the math experts. They use linear algebra and multivariable calculus to create new insight from existing data.

How exactly does this insight look? Here’s an example:

An industrial company produces a lot of products that need to be tested before shipping.

Usually such tests take a lot of time because there are hundreds of things to be tested. All to make sure that your product is not broken.

Wouldn’t it be great to know early if a test fails ten steps down the line? If you knew that you could skip the other tests and just trash the product or repair it.

That’s exactly where a data scientist can help you, big-time. This field is called predictive analytics and the technique of choice is machine learning.

Machine what? Learning?

Yes, machine learning, it works like this:

You feed an algorithm with measurement data. It generates a model and optimises it based on the data you fed it with. That model basically represents a pattern of how your data is looking You show that model new data and the model will tell you if the data still represents the data you have trained it with. This technique can also be used for predicting machine failure in advance with machine learning. Of course the whole process is not that simple.

The actual process of training and applying a model is not that hard. A lot of work for the data scientist is to figure out how to pre-process the data that gets fed to the algorithms.

Because to train a algorithm you need useful data. If you use any data for the training the produced model will be very unreliable.

A unreliable model for predicting machine failure would tell you that your machine is damaged even if it is not. Or even worse: It would tell you the machine is ok even when there is an malfunction.

Model outputs are very abstract. You also need to post-process the model outputs to receive health values from 0 to 100.

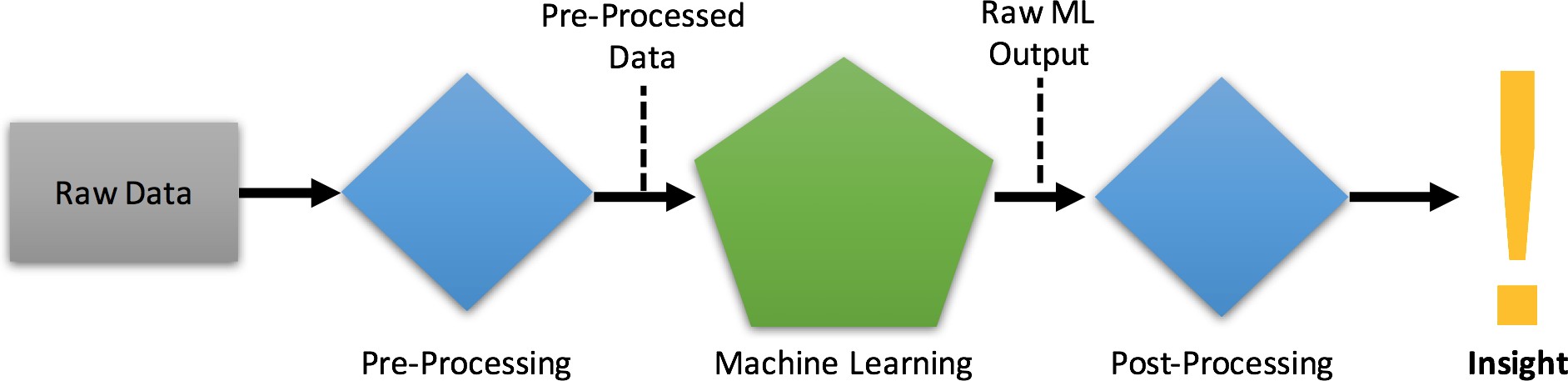


Figure 2.1: The Machine Learning Pipeline

## Data Engineer

Data Engineers are the link between the management’s big data strategy and the data scientists that need to work with data.

What they do is building the platforms that enable data scientists to do their magic. These platforms are usually used in five different ways:

* + - Data ingestion and storage of large amounts of data
    - Algorithm creation by data scientists
    - Automation of the data scientist’s machine learning models and algorithms for production use
    - Data visualisation for employees and customers
    - Most of the time these guys start as traditional solution architects for systems that involve SQL databases, web servers, SAP installations and other “standard” systems.

But to create big data platforms the engineer needs to be an expert in specifying, set- ting up and maintaining big data technologies like: Hadoop, Spark, HBase, Cassandra, MongoDB, Kafka, Redis and more.

What they also need is experience on how to deploy systems on cloud infrastructure like at Amazon or Google or on premise hardware.

**Podcast Episode:** #048 From Wannabe Data Scientist To Engineer My Journey

In this episode Kate Strachnyi interviews me for her humans of data science podcast. We talk about how I found out that I am more into the engineering part of data science.

YouTube [Click here to watch](https://youtu.be/pIZkTuN5AMM) Audio [Click here to listen](https://anchor.fm/andreaskayy/episodes/048-From-Wannabe-Data-Scientist-To-Engineer-My-Journey-e45i2o)

Table 2.2: Podcast: 048 From Wannabe Data Scientist To Engineer My Journey

## Who Companies Need

For a good company it is absolutely important to get well trained data engineers and data scientists. Think of the data scientist as the professional race car driver. A fit athlete with talent and driving skills like you have never seen.

What he needs to win races is someone who will provide him the perfect race car to drive. That’s what the solution architect is for.

Like the driver and his team the data scientist and the data engineer need to work closely together. They need to know the different big data tools Inside and out.

That’s why companies are looking for people with Spark experience. It is a common ground between both that drives innovation.

Spark gives data scientists the tools to do analytics and helps engineers to bring the data scientist’s algorithms into production. After all, those two decide how good the data

platform is, how good the analytics insight is and how fast the whole system gets into a production ready state.

# Part II

**Basic Data Engineering Skills**

# The Cloud

## IaaS vs PaaS vs SaaS

Check out this Podcast it will help you understand where’s the difference and how to decide on what you are going to use.

**Podcast Episode:** #082 Reading Tweets With Apache Nifi & IaaS vs PaaS vs SaaS

In this episode we are talking about the differences between infrastructure as a service, platform as a service and application as a service. Then we install the Nifi docker container and look into how we can extract the twitter data.

Youtube [Click here to watch](https://youtu.be/pWuT4UAocUY) Audio [Click here to listen](https://anchor.fm/andreaskayy/episodes/082-Reading-Tweets-With-Apache-Nifi--IaaS-vs-PaaS-vs-SaaS-e45j50)

Table 10.1: Podcast: 082 Reading Tweets With Apache Nifi & IaaS vs PaaS vs SaaS

## AWS,Azure, IBM, Google Cloud basics

## Cloud vs On-Premises

**Podcast Episode:** #076 Cloud vs On-Premise

How do you choose between Cloud vs On-Premises, pros and cons and what you have to think about. Because there are good reasons to not go cloud. Also thoughts on how to choose between the cloud providers by just comparing instance prices. Otherwise the comparison will drive you insane. My suggestion: Basically use them as IaaS and something like Cloudera as PaaS. Then build your solution on top of that.

YouTube [Click here to watch](https://youtu.be/BAzj0yGcrnE) Audio [Click here to listen](https://anchor.fm/andreaskayy/episodes/076-Cloud-vs-On-Premise-How-To-Decide-e45ivk)

Table 10.2: Podcast: 076 Cloud vs On-Premise

## Security

Listen to a few thoughts about the cloud in this podcast: [https://anchor.fm/andreaskayy/](https://anchor.fm/andreaskayy/embed/episodes/Dont-Be-Arrogant-The-Cloud-is-Safer-Then-Your-On-Premise-e16k9s) [embed/episodes/Dont-Be-Arrogant-The-Cloud-is-Safer-Then-Your-On-Premise-e16k9s](https://anchor.fm/andreaskayy/embed/episodes/Dont-Be-Arrogant-The-Cloud-is-Safer-Then-Your-On-Premise-e16k9s)

## Hybrid Clouds

Hybrid clouds are a mixture of on-premises and cloud deployment. A very interesting example for this is Google Anthos:

<https://cloud.google.com/anthos/>

# Security Zone Design

## How to secure a multi layered application

(UI in different zone then SQL DB)

## Cluster security with Kerberos

I talked about security zone design and lambda architecture in this podcast: [https:// anchor.fm/andreaskayy/embed/episodes/How-to-Design-Security-Zones-and-Lambda-Architecture--P](https://anchor.fm/andreaskayy/embed/episodes/How-to-Design-Security-Zones-and-Lambda-Architecture--PoDS-032-e248q2)

## Kerberos Tickets

# Big Data

## What is big data and where is the difference to data science and data analysis?

Most of you knows big data is evolving technology to process data faster. Nonetheless, big data is everywhere now, so professionals are getting tired of all this waiting and more you have

to be master the time management to plan your work effectively and not to waste the money of

your employee. Generally, as a human being we have been collecting data at exponential rate and

its only going to increase and continue to growth. Big data is a large volume of data consist of structured, semi structure and unstructured that inundates a business on a day to day basis. Its not

the amount of data likely important, the truth is what organizations do with the data.

Analytic is to find patterns in data. The ability to divide vast amount of data into logical groups and

And the ability to find hidden components

## The 4Vs of Big Data — available

It is a complete misconception. Volume is only one part of the often called four V’s of big data: Volume, velocity, variety and veracity.

**Volume** is about the size. How much data you have.

**Velocity** is about the speed of how fast the data is getting to you.

How much data is in a specific time needs to get processed or is coming into the system. This is where the whole concept of streaming data and real-time processing comes in to play.

**Variety** is the third one. It means, that the data is very different. That you have very different types of data structures.

Like CSV files, PDFs that you have stuff in XML. That you have JSON logfiles, or that you have data in some kind of a key value store.

It’s about the variety of data types from different sources that you basically want to join together. All to make an analysis based on that data.

**Veracity** is fourth and this is a very very difficult one. The issue with big data is, that it is very unreliable.

You cannot really trust the data. Especially when you’re coming from the IoT, the

Internet of Things side. Devices use sensors for measurement of temperature, pressure, acceleration and so on.

You cannot always be hundred percent sure that the actual measurement is right.

When you have data that is from for instance SAP and it contains data that is created by hand you also have problems. As you know we humans are bad at inputting stuff.

Everybody articulates different. We make mistakes, down to the spelling and that can be a very difficult issue for analytics.

I talked about the 4Vs in this podcast: https://anchor.fm/andreaskayy/embed/episodes/4- Vs-Of-Big-Data-Are-Enough-e1h2ra

### Planning is Everything

It’s all happens very very fast and you cannot react anymore. There’s a necessary type of planning and analyzing the potential of your business case necessary.

Then you need to decide if you actually have big data or not.

You need to decide if you use big data tools. This means when you conceptualize the whole infrastructure it might look ridiculous to actually focus on big data tools.

But in the long run it will help you a lot. Good planning will get a lot of problems out of the way, especially if you think about streaming data and real-time analytics.

### The Problem With ETL

A typical old-school platform deployment would look like the picture below. Devices use a data API to upload data that gets stored in a SQL database. An external analytics tool is querying data and uploading the results back to the SQL db. Users then use the user interface to display data stored in the database.

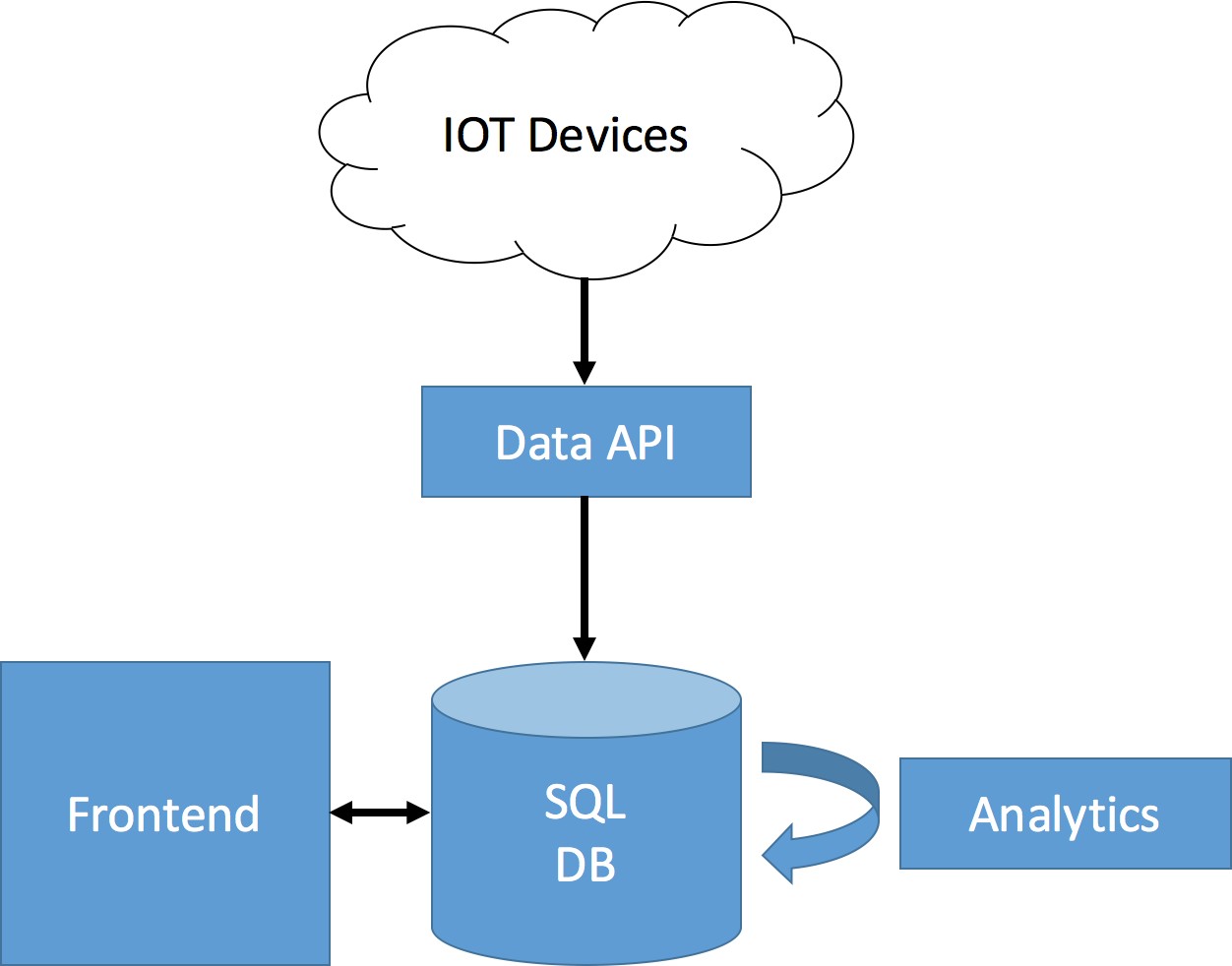


Figure 12.1: Common SQL Platform Architecture

Now, when the front end queries data from the SQL database the following three steps happen:

The database extracts all the needed rows from the storage. Extracted data gets trans- formed, for instance sorted by timestamp or something a lot more complex.

The extracted and transformed data gets loaded to the destination (the user interface) for chart creation With exploding amounts of stored data the ETL process starts being a real problem.

Analytics is working with large data sets, for instance whole days, weeks, months or more. Data sets are very big like 100GB or Terabytes. That means Billions or Trillions of rows.

This has the result that the ETL process for large data sets takes longer and longer. Very quickly the ETL performance gets so bad it won’t deliver results to analytics anymore.

A traditional solution to overcome these performance issues is trying to increase the performance of the database server. That’s what’s called scaling up.

### Scaling Up

To scale up the system and therefore increase ETL speeds administrators resort to more powerful hardware by:

Speeding up the extract performance by adding faster disks to physically read the data faster. Increasing RAM for row caching. What is already in memory does not have to be read by slow disk drives. Using more powerful CPU’s for better transform performance (more RAM helps here as well) Increasing or optimising networking performance for faster data delivery to the front end and analytics Scaling up the system is fairly easy.

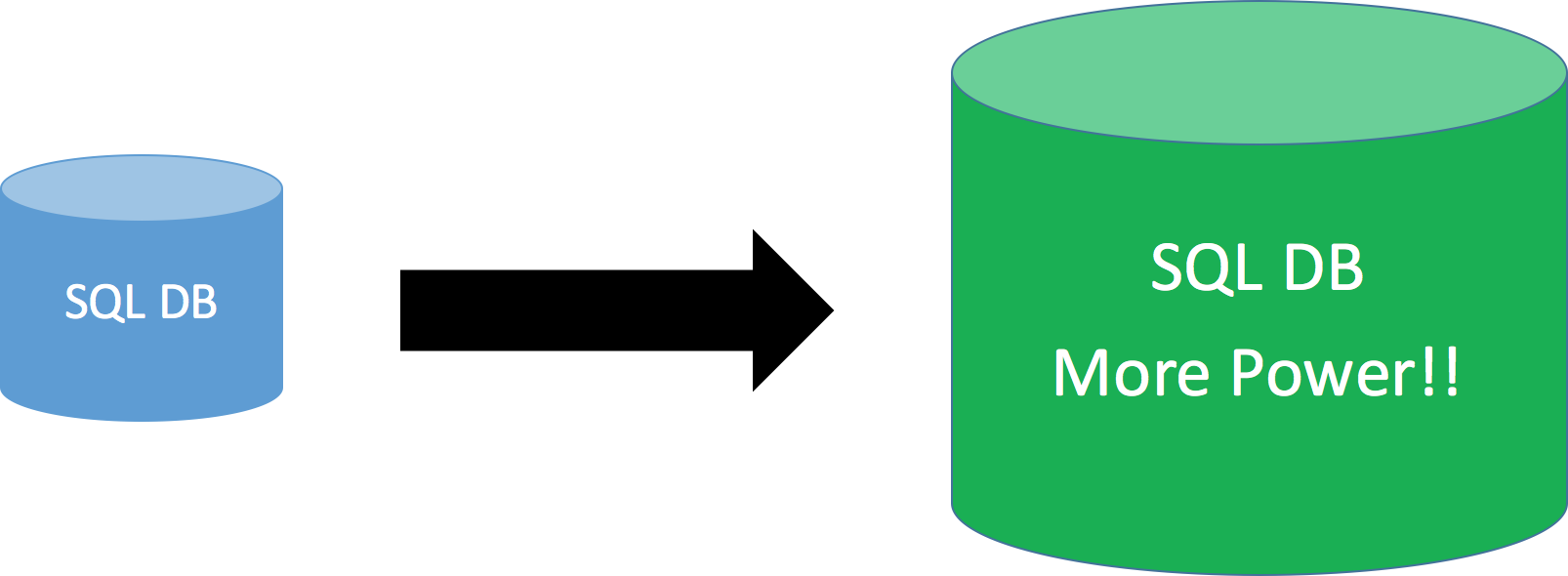


Figure 12.2: Scaling up a SQL Database

But with exponential growth it is obvious that sooner or later (more sooner than later) you will run into the same problems again. At some point you simply cannot scale up anymore because you already have a monster system, or you cannot afford to buy more expensive hardware.

The next step you could take would be scaling out.

### Scaling Out

Scaling out is the opposite of scaling up. Instead of building bigger systems the goal is to distribute the load between many smaller systems.

The simplest way of scaling out an SQL database is using a storage area network (SAN) to store the data. You can then use up to eight SQL servers, attach them to the SAN and let them handle queries. This way load gets distributed between those eight servers.

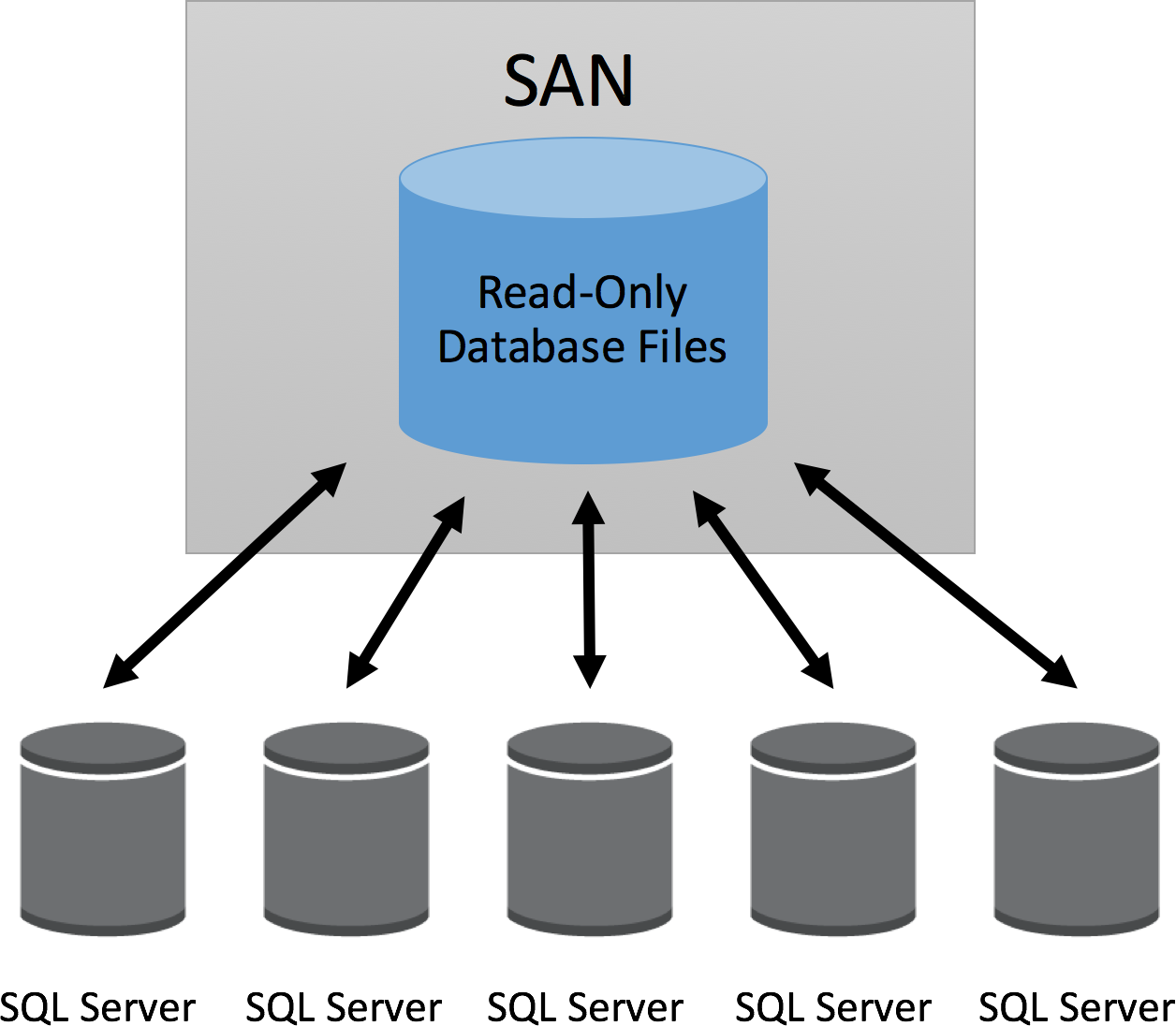


Figure 12.3: Scaling out a SQL Database

One major downside of this setup is that, because the storage is shared between the sql servers, it can only be used as an read only database. Updates have to be done periodically, for instance once a day. To do updates all SQL servers have to detach from the database. Then, one is attaching the db in read-write mode and refreshing the data. This procedure can take a while if a lot of data needs to be uploaded.

This Link to a Microsoft MSDN page has more options of scaling out an SQL database for you.

I deliberately don’t want to get into details about possible scaling out solutions. The point I am trying to make is that while it is possible to scale out SQL databases it is very complicated.

There is no perfect solution. Every option has its up- and downsides. One common major issue is the administrative effort that you need to take to implement and maintain

a scaled out solution.

### Please Don’t go Big Data

If you don’t run into scaling issues please, do not use big data tools!

Big data is a expensive thing. A Hadoop cluster for instance needs at least five servers to work properly. More is better.

Believe me this stuff costs a lot of money.

Especially when you are talking the maintenance and development on top big data tools into account.

If you don’t need it it’s making absolutely no sense at all!

On the other side: If you really need big data tools they will save your ass :)

# My Big Data Platform Blueprint

Some time ago I have created a simple and modular big data platform blueprint for myself. It is based on what I have seen in the field and read in tech blogs all over the internet.

Today I am going to share it with you.

Why do I believe it will be super useful to you?

Because, unlike other blueprints it is not focused on technology. It is based on four common big data platform design patterns.

Following my blueprint will allow you to create the big data platform that fits exactly your needs. Building the perfect platform will allow data scientists to discover new insights.

It will enable you to perfectly handle big data and allow you to make data driven decisions.

**THE BLUEPRINT** The blueprint is focused on the four key areas: Ingest, store, analyse and display.

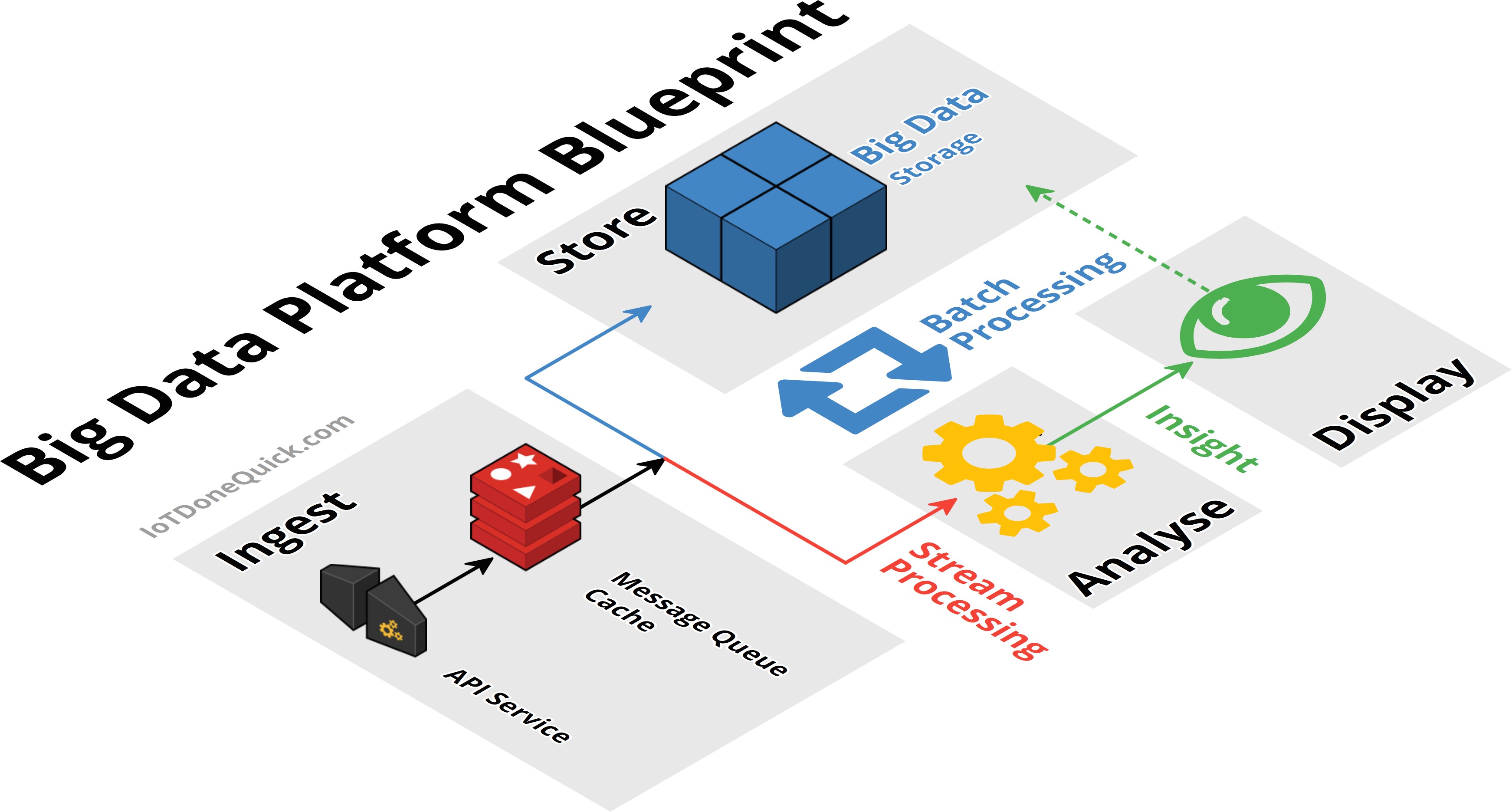


Figure 13.1: Platfrom Blueprint

Having the platform split like this turns it it a modular platform with loosely coupled interfaces.

Why is it so important to have a modular platform?

If you have a platform that is not modular you end up with something that is fixed or hard to modify. This means you can not adjust the platform to changing requirements of the company.

Because of modularity it is possible to switch out every component, if you need it. Now, lets talk more about each key area.

## Ingest

Ingestion is all about getting the data in from the source and making it available to later stages. Sources can be everything form tweets, server logs to IoT sensor data like from cars.

Sources send data to your API Services. The API is going to push the data into a temporary storage.

The temporary storage allows other stages simple and fast access to incoming data.

A great solution is to use messaging queue systems like Apache Kafka, RabbitMQ or AWS Kinesis. Sometimes people also use caches for specialised applications like Redis.

A good practice is that the temporary storage follows the publish, subscribe pattern. This way APIs can publish messages and Analytics can quickly consume them.

## Analyse / Process

The analyse stage is where the actual analytics is done. Analytics, in the form of stream and batch processing.

Streaming data is taken from ingest and fed into analytics. Streaming analyses the “live” data thus, so generates fast results.

As the central and most important stage, analytics also has access to the big data storage. Because of that connection, analytics can take a big chunk of data and analyse it.

This type of analysis is called batch processing. It will deliver you answers for the big questions.

To learn more about stream and batch processing read my blog post: How to Create New and Exciting Big Data Aided Products

The analytics process, batch or streaming, is not a one way process. Analytics also can write data back to the big data storage.

Often times writing data back to the storage makes sense. It allows you to combine previous analytics outputs with the raw data.

Analytics insight can give meaning to the raw data when you combine them. This combination will often times allow you to create even more useful insight.

A wide variety of analytics tools are available. Ranging from MapReduce or AWS Elastic MapReduce to Apache Spark and AWS lambda.

## Store

This is the typical big data storage where you just store everything. It enables you to analyse the big picture.

Most of the data might seem useless for now, but it is of upmost importance to keep it. Throwing data away is a big no no.

Why not throw something away when it is useless?

Although it seems useless for now, data scientists can work with the data. They might find new ways to analyse the data and generate valuable insight from it.

What kind of systems can be used to store big data?

Systems like Hadoop HDFS, Hbase, Amazon S3 or DynamoDB are a perfect fit to store big data.

Check out my podcast how to decide between SQL and NoSQL: https://anchor.fm/andreaskayy/embed/ Vs-SQL-How-To-Choose-e12f1o

## Display

Displaying data is as important as ingesting, storing and analysing it. People need to be able to make data driven decisions.

This is why it is important to have a good visual presentation of the data. Sometimes you have a lot of different use cases or projects using the platform.

It might not be possible for you to build the perfect UI that fits everyone. What you should do in this case is enable others to build the perfect UI themselves.

How to do that? By creating APIs to access the data and making them available to developers.

Either way, UI or API the trick is to give the display stage direct access to the data in the big data cluster. This kind of access will allow the developers to use analytics results as well as raw data to build the the perfect application.

# Lambda Architecture

**Podcast Episode:** #077 Lambda Architecture and Kappa Architecture

In this stream we talk about the lambda architecture with stream and batch process- ing as well as a alternative the Kappa Architecture that consists only of streaming. Also Data engineer vs data scientist and we discuss Andrew Ng’s AI Transformation Playbook

Audio [Click here to listen](https://anchor.fm/andreaskayy/episodes/077-Lambda--Kappa-Architecture-e45j0r) Youtube [Click here to watch](https://youtu.be/iUOQPyHN9-0)

Table 14.1: Podcast: 077 Lambda Architecture and Kappa Architecture

## Batch Processing

Ask the big questions. Remember your last yearly tax statement?

You break out the folders. You run around the house searching for the receipts. All that fun stuff.

When you finally found everything you fill out the form and send it on its way. Doing the tax statement is a prime example of a batch process.

Data comes in and gets stored, analytics loads the data from storage and creates an output (insight):

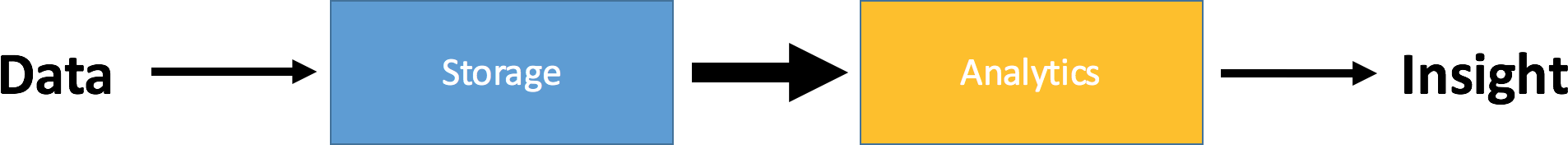


Figure 14.1: Batch Processing Pipeline

Batch processing is something you do either without a schedule or on a schedule (tax statement). It is used to ask the big questions and gain the insights by looking at the big picture.

To do so, batch processing jobs use large amounts of data. This data is provided by storage systems like Hadoop HDFS.

They can store lots of data (petabytes) without a problem.

Results from batch jobs are very useful, but the execution time is high. Because the amount of used data is high.

It can take minutes or sometimes hours until you get your results.

## Stream Processing

Gain instant insight into your data.

Streaming allows users to make quick decisions and take actions based on “real-time” insight. Contrary to batch processing, streaming processes data on the fly, as it comes in.

With streaming you don’t have to wait minutes or hours to get results. You gain instant insight into your data.

In the batch processing pipeline, the analytics was after the data storage. It had access to all the available data.

Stream processing creates insight before the data storage. It has only access to fragments of data as it comes in.

As a result the scope of the produced insight is also limited. Because the big picture is missing.

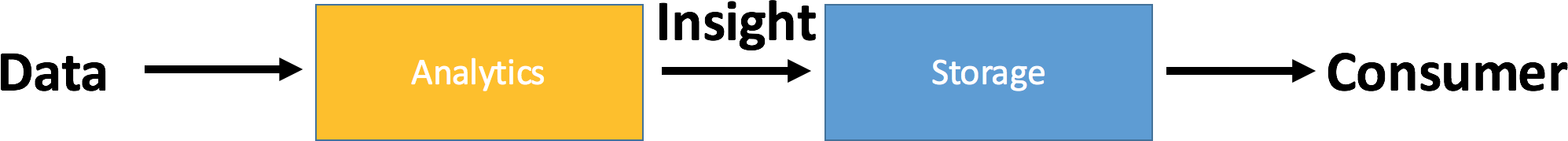


Figure 14.2: Stream Processing Pipeline

Only with streaming analytics you are able to create advanced services for the customer. Netflix for instance incorporated stream processing into Chuckwa V2.0 and the new Keystone pipeline.

One example of advanced services through stream processing is the Netflix “Trending Now” feature. Check out the Netflix case study.

## Should you do stream or batch processing?

It is a good idea to start with batch processing. Batch processing is the foundation of every good big data platform.

A batch processing architecture is simple, and therefore quick to set up. Platform sim- plicity means, it will also be relatively cheap to run.

A batch processing platform will enable you to quickly ask the big questions. They will give you invaluable insight into your data and customers.

When the time comes and you also need to do analytics on the fly, then add a streaming pipeline to your batch processing big data platform.

## Lambda Architecture Alternative

### Kappa Architecture

* + 1. **Kappa Architecture with Kudu**

## Why a Good Data Platform Is Important

**Podcast Episode:** #066 How To Do Data Science From A Data Engineers Perspective

A simple introduction how to do data science in the context of the internet of things.

YouTube [Click here to watch](https://youtu.be/yp_cc4R0mGQ) Audio [Click here to listen](https://anchor.fm/andreaskayy/episodes/066-How-To-Do-Data-Science-From-A-Data-Engineers-Perspective-e45imt)

Table 14.2: Podcast: 066 How To Do Data Science From A Data Engineers..

# Data Warehouse vs Data Lake

**Podcast Episode:** #055 Data Warehouse vs Data Lake

On this podcast we are going to talk about data warehouses and data lakes? When do people use which? What are the pros and cons of both? Architecture examples for both Does it make sense to completely move to a data lake?

YouTube [Click here to watch](https://youtu.be/8gNQTrUUwMk) Audio [Click here to listen](https://anchor.fm/andreaskayy/episodes/055-Data-Warehouse-vs-Data-Lake-e45iem)

Table 15.1: Podcast: 055 Data Warehouse vs Data Lake

A picture containing screenshot

Description automatically generated

**Data warehouse** is something very traditional and decades where you get in data in structured form, filtered data

that has already been processed for a specific purpose where the data coming from production machine data into data warehouse.

On the left side data source layer which consist of marketing, sales and inventory and social media data. The idea

Is you already have those data sources and want to get in very structured manner so it can be present and analyze.

Data warehouse have a staging layer where the data less more put into and more likely a temporary storage. Once the extracted has been loaded, it will be subjected to high-level data quality checks. The final result will be clean and organized data that you will load into your data warehouse

Extract, Transform and Load tools (ETL) are the data integration tools used to extract data from source systems, transform and prepare data and load into the data warehouse(database). Panoply.io product provides this entire process, easily and quickly. All you need to do is point it to your data source(s).

And from the storage layer it goes to Presentation layer where it actually do the reporting and analysis.

**4 key differences between data warehouse and data lake**

**Data lake**

* Very raw structured and do not have a lot of preprocessed data. Pull the data and work it later.

**Data warehouse**

* Something that very preprocess and clearly defined

**Data lake** is a vast pool of raw data(unstructured), the purpose for which is not yet defined.

# Hadoop Platforms — available

When people talk about big data, one of the first things come to mind is Hadoop. Google’s search for Hadoop returns about 28 million results. Industry has realized that we need cluster of computers to process big amount of data as the data to be processed was too much for the capacity of single powerful servers.

It seems like you need Hadoop to do big data. Today I am going to shed light onto why Hadoop is so trendy.

You will see that Hadoop has evolved from a platform into an ecosystem. It’s design allows a lot of Apache projects and 3rd party tools to benefit from Hadoop.

## What is Hadoop

Hadoop is a platform for distributed storing and analyzing of very large data sets, that is affordable

And easy to obtain.

Hadoop has four main modules: Hadoop common, HDFS, MapReduce and YARN. The way these modules are woven together is what makes Hadoop so successful.

The Hadoop common libraries and functions are working in the background. That’s why I will not go further into them. They are mainly there to support Hadoop’s modules.

**Podcast Episode:** #060 What Is Hadoop And Is Hadoop Still Relevant In 2019?

A Introduction into Hadoop HDFS, YARN and MapReduce. Yes, Hadoop is still relevant in 2019 even if you look into serverless tools.

YouTube [Click here to watch](https://youtu.be/8AWaht3YQgo) Audio [Click here to listen](https://anchor.fm/andreaskayy/episodes/060-What-Is-Hadoop-And-Is-Hadoop-Still-Relevant-In-2019-e45ijp)

Table 16.1: Podcast: 060 What Is Hadoop And Is Hadoop Still Relevant In 2019?

## Hadoop main components

## Common terms used for Hadoop cluster

CLUSTER – A group of inter-connected computers that can share processing power and storage

NODE – A single node is an individual entity handling certain responsibilities in a cluster (Master and Worker/Slave)

MASTER NODE – The node that handles management of resources, assignment and killing of tasks, and monitoring of job completion

WORKER/SLAVE NODE – Responsible in running the jobs assigned by master node

DAEMON – An instance of task running on a node

## What makes Hadoop so popular? — available

Storing and analyzing data as large as you want is nice. But what makes Hadoop so popular?

Hadoops core functionality is the driver of Hadoop adoption.Many Apache side projects use its core functions.

Because of all those side projects Hadoop has turned more into an ecosystem. An ecosys- tem for storing and processing big data.

To better visualize this eco system I have drawn you the following graphic. It shows some projects of the Hadoop ecosystem who are closely connected with the Hadoop.

It is not a complete list. There are many more tools that even I don’t know. Maybe I am drawing a complete map in the future.

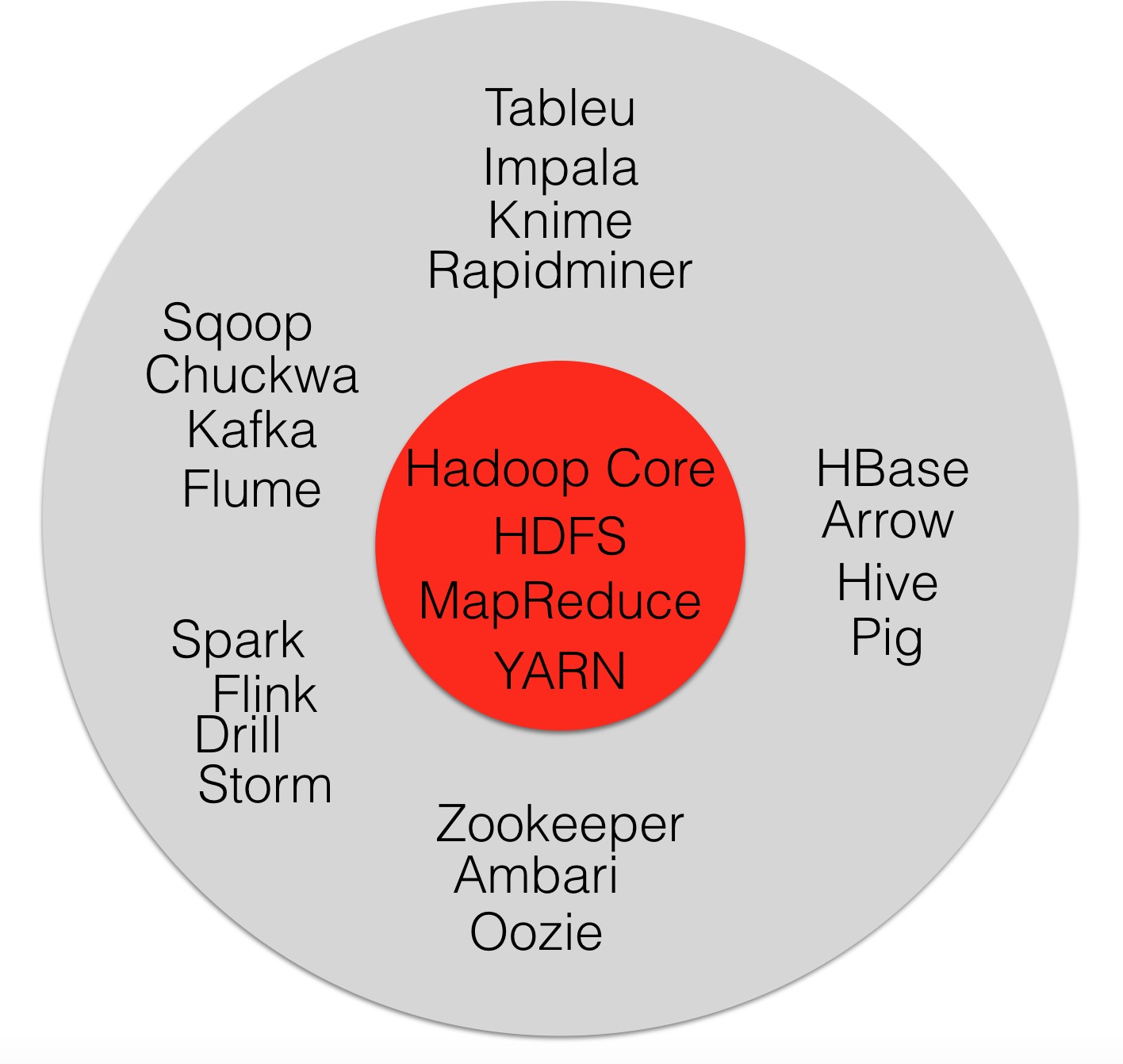


Figure 16.1: Hadoop Ecosystem Components

## Hadoop Ecosystem Components

Remember my big data platform blueprint? The blueprint has four stages: Ingest, store, analyse and display.

Because of the Hadoop ecosystem” the different tools in these stages can work together perfectly.

Here’s an example:

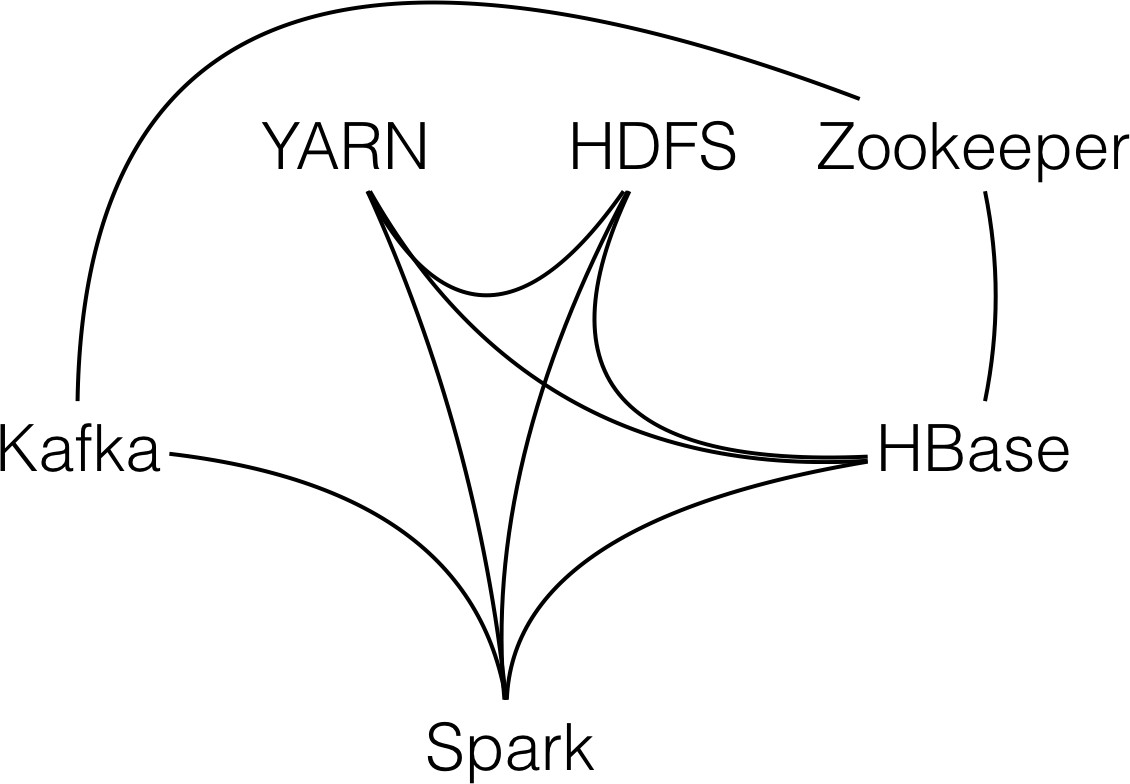


Figure 16.2: Connections between tools

You use Apache Kafka to ingest data, and store the it in HDFS. You do the analytics with Apache Spark and as a backend for the display you store data in Apache HBase.

To have a working system you also need YARN for resource management. You also need Zookeeper, a configuration management service to use Kafka and HBase

As you can see in the picture below each project is closely connected to the other.

Spark for instance, can directly access Kafka to consume messages. It is able to access HDFS for storing or processing stored data.

It also can write into HBase to push analytics results to the front end.

The cool thing of such ecosystem is that it is easy to build in new functions. Want to store data from Kafka directly into HDFS without using Spark?

No problem, there is a project for that. Apache Flume has interfaces for Kafka and HDFS.

It can act as an agent to consume messages from Kafka and store them into HDFS. You even do not have to worry about Flume resource management.

Flume can use Hadoop’s YARN resource manager out of the box.

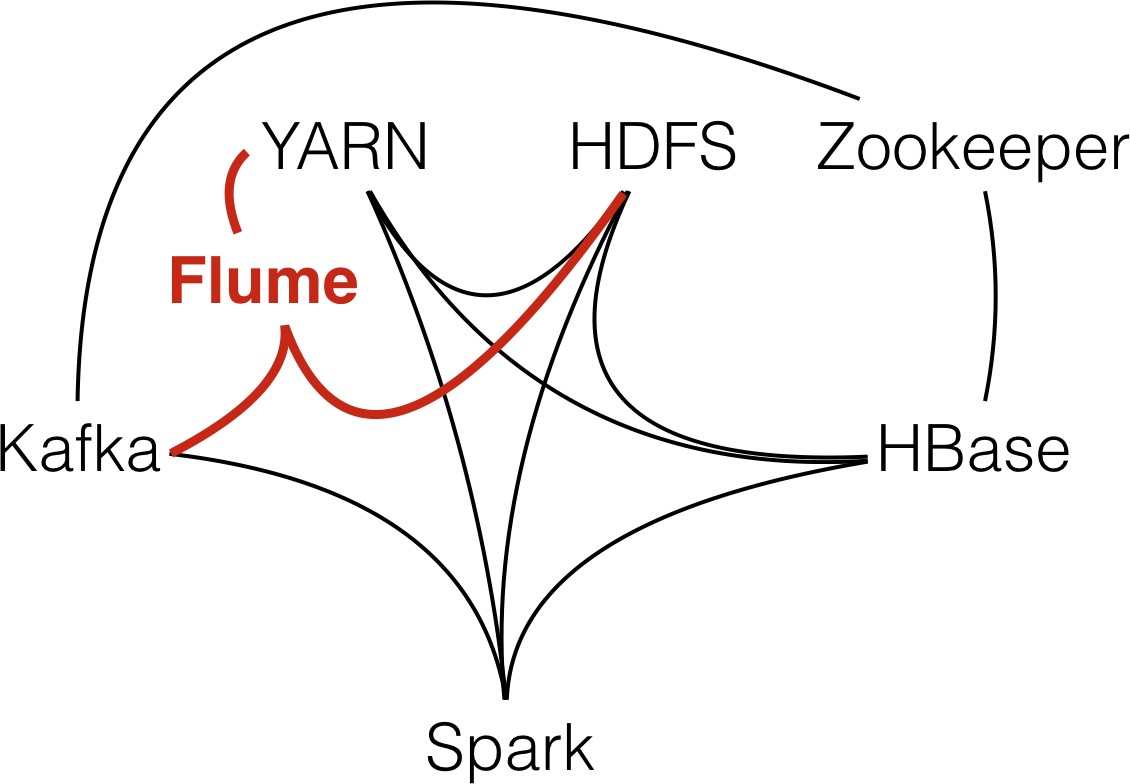


Figure 16.3: Flume Integration

## Hadoop Is Everywhere?

Although Hadoop is so popular it is not the silver bullet. It isn’t the tool that you should use for everything.

Often times it does not make sense to deploy a Hadoop cluster, because it can be overkill. Hadoop does not run on a single server.

You basically need at least five servers, better six to run a small cluster. Because of that. the initial platform costs are quite high.

One option you have is to use a specialized systems like Cassandra, MongoDB or other NoSQL DB’s for storage. Or you move to Amazon and use Amazon’s Simple Storage Service, or S3.

Guess what the tech behind S3 is. Yes, HDFS. That’s why AWS also has the equivalent to MapReduce named Elastic MapReduce.

The great thing about S3 is that you can start very small. When your system grows you don’t have to worry about s3’s server scaling.

## Should you learn Hadoop?

Yes, I definitely recommend you to get to now how Hadoop works and how to use it. As I have shown you in this article, the ecosystem is quite large.

Many big data projects use Hadoop or can interface with it. Thats why it is generally a good idea to know as many big data technologies as possible.

Not in depth, but to the point that you know how they work and how you can use them. Your main goal should be to be able to hit the ground running when you join a big data project.

Plus, most of the technologies are open source. You can try them out for free.

**How does a Hadoop System architecture look like** **What tools are usually in a with Hadoop Cluster** Yarn Zookeeper HDFS Oozie Flume Hive

## How to select Hadoop Cluster Hardware

# Docker

## What is docker and what do you use it for — available

Have you played around with Docker yet? If you’re a data science learner or a data scientist you need to check it out!

It’s awesome because it simplifies the way you can set up development environments for data science. If you want to set up a dev environment you usually have to install a lot of packages and tools.

### Don’t Mess Up Your System

What this does is you basically mess up your operating system. If you’re a starter you don’t know which packages you need to install. You don’t know which tools you need to install.

If you want to for instance start with Jupyter notebooks you need to install that on your PC somehow. Or you need to start installing tools like PyCharmor Anaconda.

All that gets added to your system and so you mess up your system more and more and more. What Docker brings you, especially if you’re on a Mac or a Linux system is simplicity.

### Preconfigured Images

Because it is so easy to install on those systems. Another cool thing about docker images is you can just search them in the Docker store, download them and install them on your system.

Running them in a completely pre-configured environment. You don’t need to think about stuff you go to the Docker library you search for deep learning GPU and Python.

You get a list of images you can download. You download one, start it up, you go to the browser hit up the URL and just start coding.

Start doing the work. The only other thing you need to do is bind some drives to that instance so you can exchange files. And then that’s it!

There is no way that you can crash or mess up your system. It’s all encapsulated into Docker.Why this works is because Docker has natively access to your hardware.

### Take It With You

It’s not a completely virtualized environment like a VirtualBox. An image has the upside that you can take it wherever you want. So if your on your PC at home use that there.

Make a quick build, take the image and go somewhere else. Install the image which is usually quite fast and just use it like you’re at home.

It’s that awesome!

## Kubernetes Container Deployment

I am getting into Docker a lot more myself. For a bit different reasons.

What I’m looking for is using Docker with Kubernetes. With Kubernetes you can auto- mate the whole container deployment process.

The idea with is that you have a cluster of machines. Lets say you have 10 server cluster and you run Kubernetes on them.

Kubernetes lets you spin up Docker containers on-demand to execute tasks. You can set up how much resources like CPU, RAM, Network, Docker container can use.

You can basically spin up containers, on the cluster on demand. When ever you need to do a analytics task.

Perfect for Data Science.

## How to create, start,stop a Container

## Docker micro services?

## Kubernetes

## Why and how to do Docker container orchestration

Podcast about how data science learners use Docker (for data scientists): https://anchor.fm/andreaskay Data-Science-Go-Docker-e10n7u

## Useful Docker Commands

Create a container:

docker run CONTAINER ***−*** network NETWORK

Start a stopped container:

docker s t a r t CONTAINERNAME

Stop a running container:

docker stop

List all running containers

docker ps

List all containers including stopped ones

docker ps *−*a

Inspect the container configuration. For instance network settings and so on:

docker i ns pe c t CONTAINER

List all available virtual networks:

docker network l s

Create a new network:

docker network c re ate NETWORK ***−*** dri v e r bridge

Connect a running container to a network

docker network connect NETWORK CONTAINER

Disconnect a running container from a network

docker network disconnect NETWORK CONTAINER

Remove a network

network rm NETWORK

# REST APIs

APIs, Application Programming Interfaces are the cornerstones of any great data plat- form.

**Podcast Episode:** #033 How APIs Rule The World

Strong APIs make a good platform. In this episode I talk about why you need APIs and why Twitter is a great example. Especially JSON APIs are my personal favorite. Because JSON is also important in the Big Data world, for instance in log analytics. How? Check out this episode!

Audio [Click here to listen](https://anchor.fm/andreaskayy/episodes/How-APIs-Rule-The-World--PoDS-033-e24ttq)

Table 18.1: Podcast: 033 How APIs Rule The World

## API Design

In this podcast episode we look into the Twitter API. It’s a great example how to build an API

**Podcast Episode:** #081 Twitter API Research Data Engineering Course Part 5 In this episode we look into the Twitter API documentation, which I love by the way. How can we get old tweets for a certain hashtags and how to get current live tweets for these hashtags?

Audio [Click here to listen](https://anchor.fm/andreaskayy/episodes/081-How-to-get-tweets-from-the-Twitter-API-e45j32) Youtube [Click here to watch](https://youtu.be/UnAXKxeIlyg)

Table 18.2: Podcast: 081 Twitter API Research

## Implementation Frameworks

Jersey:

<https://jersey.github.io/documentation/latest/getting-started.html> Swagger:

<https://github.com/swagger-api/swagger-core/wiki/Swagger-2.X---Getting-started> Jersey vs Swagger:

<https://stackoverflow.com/questions/36997865/what-is-the-difference-between-swagger-api-and-jax-rs> Spring Framework:

<https://spring.io/>

When to use Spring or Jersey:

[https://stackoverflow.com/questions/26824423/what-is-the-difference-among-spring-rest-service-and-j](https://stackoverflow.com/questions/26824423/what-is-the-difference-among-spring-rest-service-and-jersey-rest-service-and-spr)

## OAuth security

# Databases

## SQL Databases

### PostgreSQL DB

Homepage:

<https://www.postgresql.org/> Postgre vs Mongodb:

<https://blog.panoply.io/postgresql-vs-mongodb>

### Database Design

* + 1. **SQL Queries**
    2. **Stored Procedures**
    3. **ODBC/JDBC Server Connections**

## NoSQL Stores

* + 1. **KeyValue Stores (HBase)**

### Document Store HDFS — available

The Hadoop distributed file system, or HDFS, allows you to store files in Hadoop. The difference between HDFS and other file systems like NTFS or EXT is that it is a dis- tributed one.

**Podcast Episode:** #056 NoSQL Key Value Stores Explained with HBase

What is the difference between SQL and NoSQL? In this episode I show you on the example of HBase how a key/value store works.

YouTube [Click here to watch](https://youtu.be/67hIkbpzFc8) Audio [Click here to listen](https://anchor.fm/andreaskayy/episodes/056-NoSQL-Key-Value-Stores-Explained-With-HBase-e45ifb)

Table 19.1: Podcast: 056 NoSQL Key Value Stores Explained with HBase

What does that mean exactly?

A typical file system stores your data on the actual hard drive. It is hardware dependent.

If you have two disks then you need to format every disk with its own file system. They are completely separate.

You then decide on which disk you physically store your data.

HDFS works different to a typical file system. HDFS is hardware independent. Not only does it span over many disks in a server. It also spans over many servers.

HDFS will automatically place your files somewhere in the Hadoop server collective.

It will not only store your file, Hadoop will also replicate it two or three times (you can define that). Replication means replicas of the file will be distributed to different servers.

This gives you superior fault tolerance. If one server goes down, then your data stays available on a different server.

Another great thing about HDFS is, that there is no limit how big the files can be. You can have server log files that are terabytes big.

How can files get so big? HDFS allows you to append data to files. Therefore, you can continuously dump data into a single file without worries.

HDFS physically stores files different then a normal file system. It splits the file into blocks.

These blocks are then distributed and replicated on the Hadoop cluster. The splitting happens automatically.

In the configuration you can define how big the blocks should be. 128 megabyte or 1 gigabyte?

No problem at all.

This mechanic of splitting a large file in blocks and distributing them over the servers is great for processing. See the MapReduce section for an example.

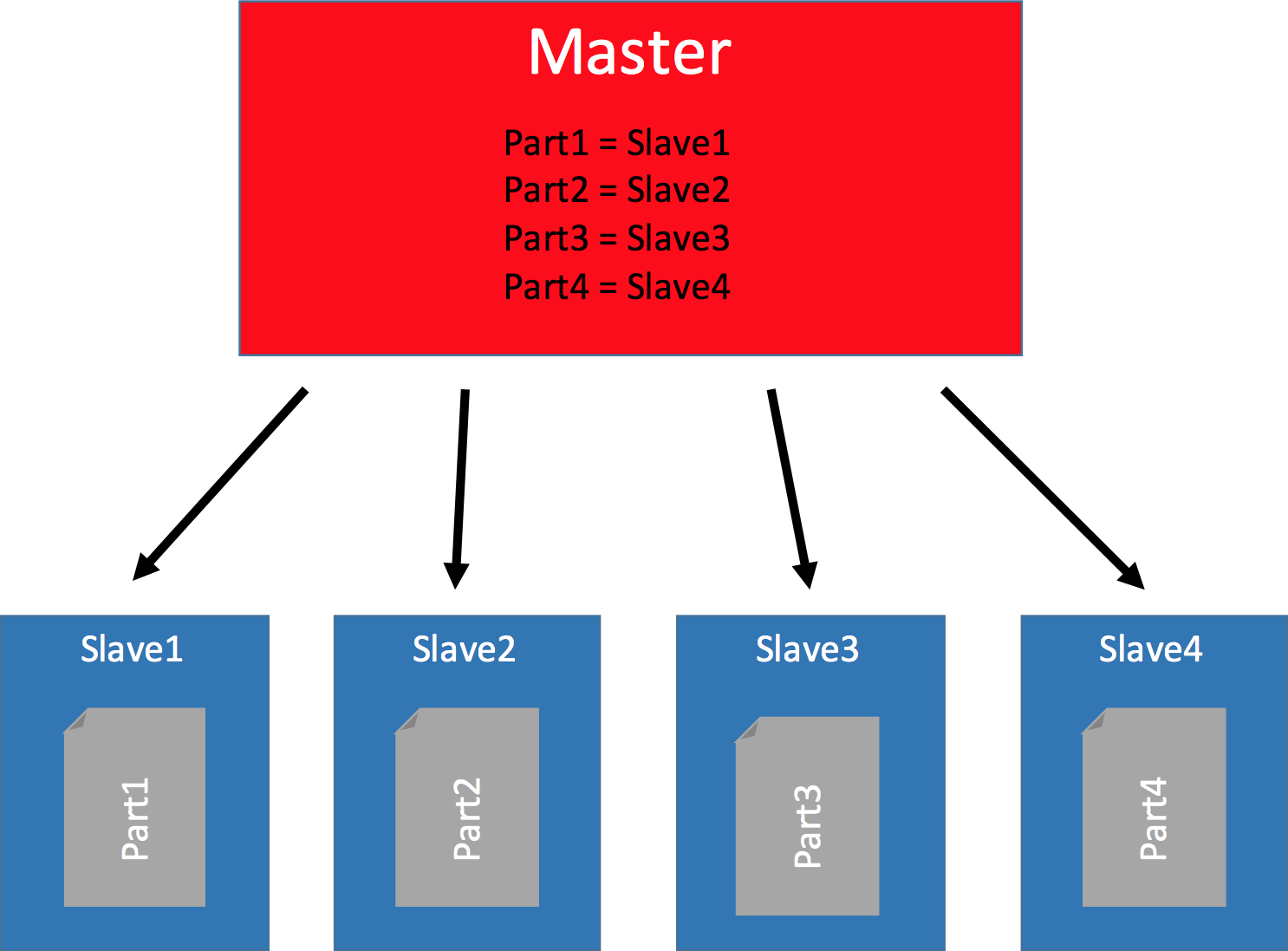


Figure 19.1: HDFS Master and Data Nodes

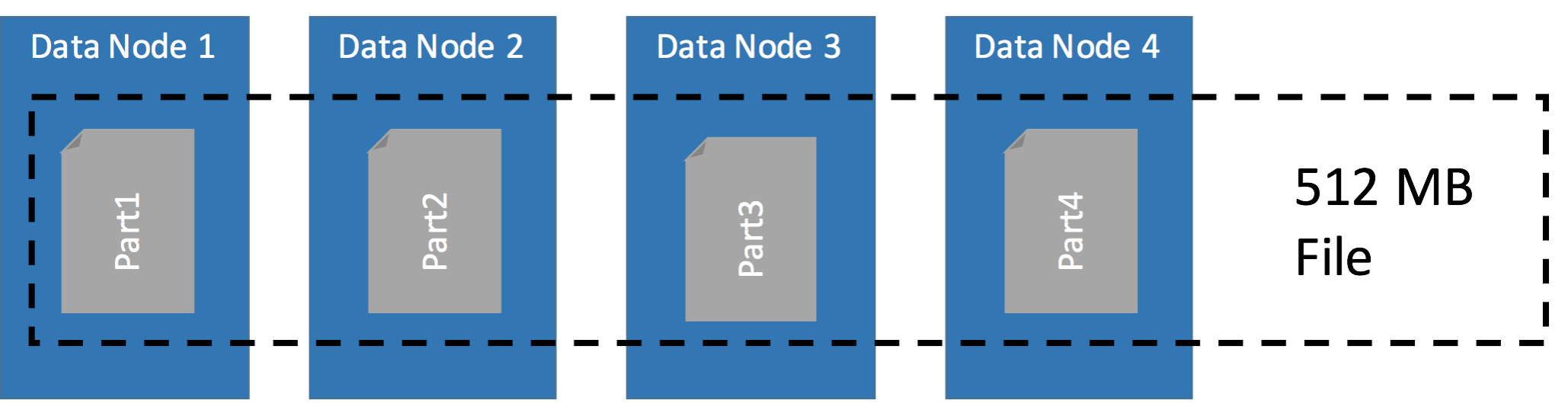


Figure 19.2: Distribution of Blocks for a 512MB File

### Document Store MongoDB

**Podcast Episode:** #093 What is MongoDB

I was always curious about MongoDB. In this stream we go over the links you can find below

YouTube [Click here to watch](https://youtu.be/U05knQN29FA)

Table 19.2: Podcast: 093 What is MongoDB

#### Links:

What is mongoDB:

<https://www.guru99.com/what-is-mongodb.html#4> Or directly from MongoDB.com: <https://www.mongodb.com/what-is-mongodb> Storage in BSON files: <https://en.wikipedia.org/wiki/BSON>

Hello World in MongoDB: <https://www.mkyong.com/mongodb/mongodb-hello-world-example> Real-Time Analytics on MongoDB Data in Power BI:

<https://dzone.com/articles/real-time-analytics-on-mongodb-data-in-power-bi> Spark and Mongodb:

<https://www.mongodb.com/scale/when-to-use-apache-spark-with-mongodb> MongoDB vs Time Series Database:

[https://blog.timescale.com/how-to-store-time-series-data-mongodb-vs-timescaledb-postgresql-a739397](https://blog.timescale.com/how-to-store-time-series-data-mongodb-vs-timescaledb-postgresql-a73939734016/)

Fun article titled why you should never use mongodb: <http://www.sarahmei.com/blog/2013/11/11/why-you-should-never-use-mongodb/> MongoDB vs Cassandra:

<https://blog.panoply.io/cassandra-vs-mongodb>

### Elasticsearch Search Engine and Document Store

Elasticsearch is not a DB but firstly a search engine that indexes JSON documents.

**Podcast Episode:** #095 What is Elasticsearch & Why is It So Popular?

Elasticsearch is a super popular tool for indexing and searching data. On this stream we check out how it works, architectures and what to use it for. There must be a reason why it is so popular.

YouTube [Click here to watch](https://youtu.be/hNb5zB4OPXM)

Table 19.3: Podcast: What is Elasticsearch & Why is It So Popular?

Links:

Great example for architecture with Elasticsearch, Logstash and Kibana: <https://www.elastic.co/pdf/architecture-best-practices.pdf>

Introduction to Elasticsearch in the documentation: <https://www.elastic.co/guide/en/elasticsearch/reference/current/elasticsearch-intro.html>

Working with JSON documents: <https://www.slideshare.net/openthinklabs/03-elasticsearch-data-in-data-out>

JSONs need to be flattened heres how to work with nested objects in the JSON: <https://www.elastic.co/guide/en/elasticsearch/reference/current/nested.html>

Indexing basics:

<https://www.slideshare.net/knoldus/deep-dive-into-elasticsearch>

How to query data with DSL language: [https://www.elastic.co/guide/en/elasticsearch/reference/current/query-dsl-common-](https://www.elastic.co/guide/en/elasticsearch/reference/current/query-dsl-common-terms-query.html)terms-query. [html](https://www.elastic.co/guide/en/elasticsearch/reference/current/query-dsl-common-terms-query.html)

How to do searches with search API: <https://www.elastic.co/guide/en/elasticsearch/reference/current/search.html>

General recommendations when working with Elasticsearch: [https://www.elastic.co/guide/en/elasticsearch/reference/current/general-recommendat](https://www.elastic.co/guide/en/elasticsearch/reference/current/general-recommendations.html)ions. [html](https://www.elastic.co/guide/en/elasticsearch/reference/current/general-recommendations.html)

JSON document example and intro to Kibana: <https://www.slideshare.net/objectrocket/an-intro-to-elasticsearch-and-kibana>

How to connect Tableau to Elasticsearch: [https://www.elastic.co/guide/en/elasticsearch/reference/current/sql-client-apps-tableau.](https://www.elastic.co/guide/en/elasticsearch/reference/current/sql-client-apps-tableau.html) [html](https://www.elastic.co/guide/en/elasticsearch/reference/current/sql-client-apps-tableau.html)

Benchmarks how fast Elasticsearch is:

<https://medium.appbase.io/benchmarking-elasticsearch-1-million-writes-per-sec-bf37e7ca8a4c>

Elasticsearch vs MongoDB quick overview:

<https://db-engines.com/en/system/Elasticsearch%3BMongoDB>

Logstash overview (preprocesses data before insert into Elasticsearch) [https://www.elastic.](https://www.elastic.co/products/logstash) [co/products/logstash](https://www.elastic.co/products/logstash)

X-Pack Security for Elasticsearch:<https://www.elastic.co/guide/en/elasticsearch/reference/current/security-api.html>

Google Trends Grafana vs Kibana: [https://trends.google.com/trends/explore?geo=US&q=%2Fg%2F11fy132gmf,%2Fg%2F11cknd0blr](https://trends.google.com/trends/explore?geo=US&amp;q=%2Fg%2F11fy132gmf%2C%2Fg%2F11cknd0blr)

### Hive Warehouse

### Impala

### Kudu

### Apache Druid

**Podcast Episode:** Druid NoSQL DB And Analytics DB Introduction

In this video I explain what Druid is and how it works. We look into the architecture of a Druid cluster and check out how Clients access the data.

YouTube [Click here to watch](https://youtu.be/EiEIeBXSWjM)

Table 19.4: Podcast: Druid NoSQL DB And Analytics DB Introduction

### InfluxDB Time Series Database

Key concepts: [https://docs.influxdata.com/influxdb/v1.7/concepts/key concepts/](https://docs.influxdata.com/influxdb/v1.7/concepts/key_concepts/) InfluxDB and Spark Streaming

[https://towardsdatascience.com/processing-time-series-data-in-real-time-with-influxdb-and-structured](https://towardsdatascience.com/processing-time-series-data-in-real-time-with-influxdb-and-structured-streaming-d1864154cf8b)

Building a Sreaming application with spark, grafana, chronogram and influx:

[https://medium.com/@xaviergeerinck/building-a-real-time-streaming-dashboard-with-spark-grafana-c](https://medium.com/%40xaviergeerinck/building-a-real-time-streaming-dashboard-with-spark-grafana-chronograf-and-influxdb-e262b68087de)

Performance Dashboard Spark and InfluxDB:

<https://db-blog.web.cern.ch/blog/luca-canali/2019-02-performance-dashboard-apache-spark>

Other alternatives for time series databases are: DalmatinerDB, InfluxDB, Prometheus, Riak TS, OpenTSDB, KairosDB

### MPP Databases (Greenplum)

# Data Processing and Analytics - Frameworks

## Is ETL still relevant for Analytics?

**Podcast Episode:** #039 Is ETL Dead For Data Science & Big Data?

Is ETL dead in Data Science and Big Data? In today’s podcast I share with you my views on your questions regarding ETL (extract, transform, load). Is ETL still practiced or did pre processing & cleansing replace it What would replace ETL in Data Engineering

YouTube [Click here to watch](https://youtu.be/leSOWPaNkl4) Audio [Click here to listen](https://anchor.fm/andreaskayy/episodes/Is-ETL-Dead-For-Data-Science--Big-Data---PoDS-039-e2b604)

Table 20.1: Podcast: 039 Is ETL Dead for Data Science and Big Data?

## Stream Processing

### Three methods of streaming — available

In stream processing sometimes it is ok to drop messages, other times it is not. Sometimes it is fine to process a message multiple times, other times that needs to be avoided like hell.

Today’s topic are the different methods of streaming: At most once, at least once and exactly once.

What this means and why it is so important to keep them in mind when creating a solution. That is what you will find out in this article.

### At Least Once

At least once, means a message gets processed in the system once or multiple times. So with at least once it’s not possible that a message gets into the system and is not getting processed.

It’s not getting dropped or lost somewhere in the system.

One example where at least once processing can be used is when you think about a fleet management of cars. You get GPS data from cars and that GPS data is transmitted with a timestamp and the GPS coordinates.

It’s important that you get the GPS data at least once, so you know where the car is. If you’re processing this data multiple times, it always has the the timestamp with it.

Because of that it does not matter that it gets processed multiple times, because of the timestamp. Or that it would be stored multiple times, because it would just override the existing one.

### At Most Once

The second streaming method is at most once. At most once means that it’s okay to drop some information, to drop some messages.

But it’s important that a message is only only processed once as a maximum.

A example for this is event processing. Some event is happening and that event is not important enough, so it can be dropped. It doesn’t have any consequences when it gets dropped.

But when that event happens it’s important that it does not get processed multiple times. Then it would look as if the event happened five or six times instead of only one.

Think about engine misfires. If it happens once, no big deal. But if the system tells you it happens a lot you will think you have a problem with your engine.

### Exactly Once

Another thing is exactly once, this means it’s not okay to drop data, it’s not okay to lose data and it’s also not okay to process one message what data said multiple times

A example for this is for instance banking. When you think about credit card transactions it’s not okay to drop a transaction.

When dropped your payment is not going through. It’s also not okay to have a transaction processed multiple times, because then you are paying multiple times.

### Check The Tools!

All of this sounds very simple and logical. What kind of processing is done has to be a requirement for your use case.

It needs to be thought about in the design process, because not every tool is supporting all three methods. Very often you need to code your application very differently based on the streaming method.

Especially exactly once is very hard to do.

So, the tool of data processing needs to be chosen based on if you need exactly once, at least once or if you need at most once.

## MapReduce

Since the early days of the Hadoop eco system, the MapReduce framework is one of the main components of Hadoop alongside the Hadoop file system HDFS.

Google for instance used MapReduce to analyse stored html content of websites through counting all the html tags and all the words and combinations of them (for instance headlines). The output was then used to create the page ranking for Google Search.

That was when everybody started to optimise his website for the google search. Serious search engine optimisation was borne. That was the year 2004.

How MapReduce is working is, that it processes data in two phases: The map phase and the reduce phase.

In the map phase, the framework is reading data from HDFS. Each dataset is called an input record.

Then there is the reduce phase. In the reduce phase, the actual computation is done and the results are stored. The storage target can either be a database or back HDFS or something else.

After all it’s Java – so you can implement what you like.

The magic of MapReduce is how the map and reduce phase are implemented and how both phases are working together.

The map and reduce phases are parallelised. What that means is, that you have multiple map phases (mappers) and reduce phases (reducers) that can run in parallel on your cluster machines.

Here’s an example how such a map and reduce process works with data:

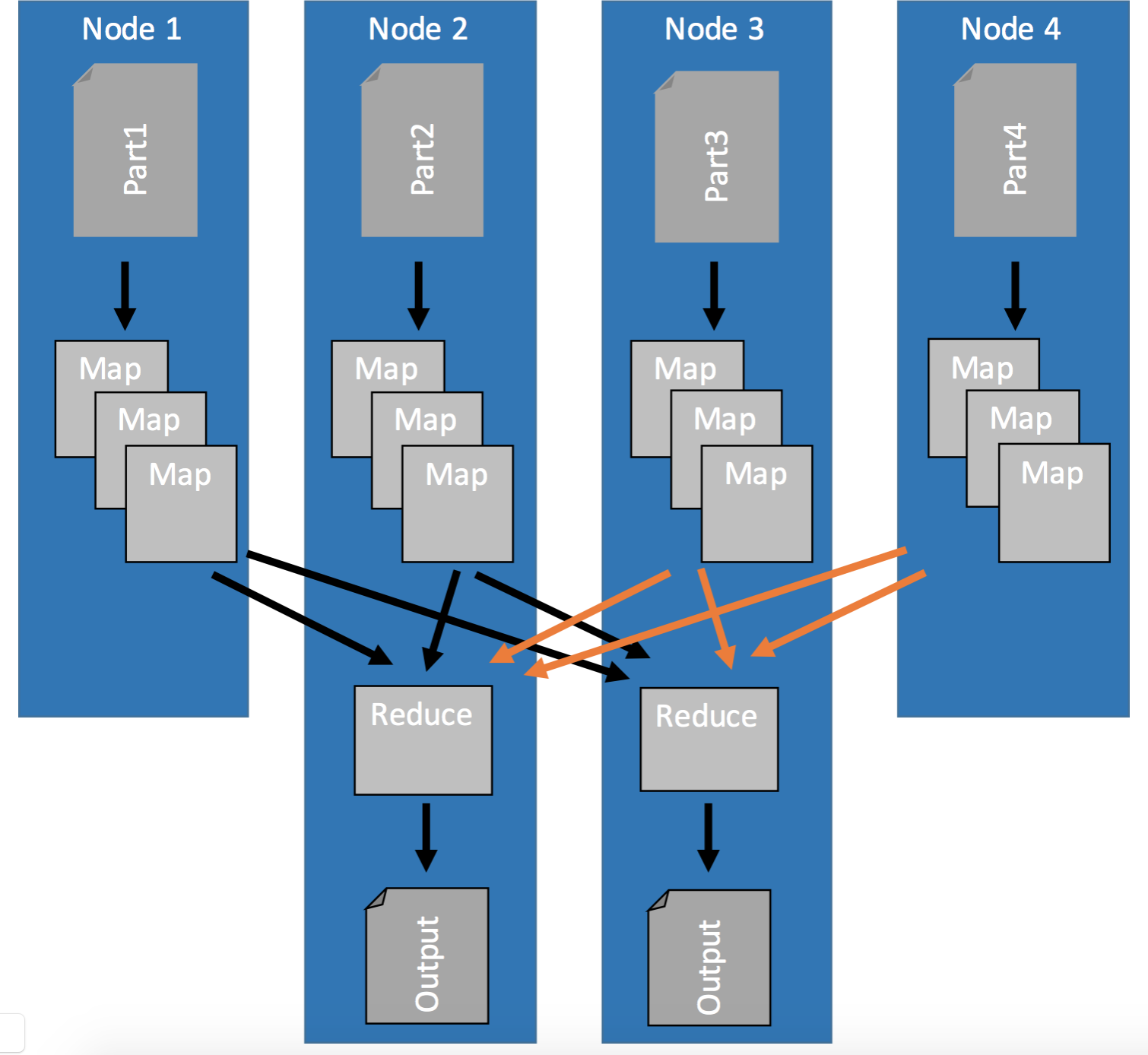


Figure 20.1: Mapping of input files and reducing of mapped records

### How does MapReduce work – available

First of all, the whole map and reduce process relies heavily on using key/value pairs. That’s what the mappers are for.

In the map phase input data, for instance a file, gets loaded and transformed into key/- value pairs.

When each map phase is done it sends the created key/value pairs to the reducers where they are getting sorted by key. This means, that an input record for the reduce phase is a list of values from the mappers that all have the same key.

Then the reduce phase is doing the computation of that key and its values and outputting the results.

How many mappers and reducers can you use in parallel? The number of parallel map and reduce processes depends on how many CPU cores you have in your cluster. Every mapper and every reducer is using one core.

This means that the more CPU cores you actually have, the more mappers you can use, the faster the extraction process can be done. The more reducers you are using the faster the actual computation is being done.

To make this more clear, I have prepared an example:

### Example

As I said before, MapReduce works in two stages, map and reduce. Often these stages are explained with a word count task.

Personally, I hate this example because counting stuff is to trivial and does not really show you what you can do with MapReduce. Therefore, we are going to use a more real world use-case from the world of the internet of things (IoT).

IoT applications create an enormous amount of data that has to be processed. This data is generated by physical sensors who take measurements, like room temperature at 8.00 o’Clock.

Every measurement consists of a key (the timestamp when the measurement has been taken) and a value (the actual value measured by the sensor).

Because you usually have more than one sensor on your machine, or connected to your system, the key has to be a compound key. Compound keys contain additionally to the measurement time information about the source of the signal.

But, let’s forget about compound keys for now. Today we have only one sensor. Each measurement outputs key/value pairs like: Timestamp-Value.

The goal of this exercise is to create average daily values of that sensor’s data.

The image below shows how the map and reduce process works.

First, the map stage loads unsorted data (input records) from the source (e.g. HDFS) by key and value (key:2016-05-01 01:02:03, value:1).

Then, because the goal is to get daily averages, the hour:minute:second information is cut from the timestamp.

That is all that happens in the map phase, nothing more.

After all parallel map phases are done, each key/value pair gets sent to the one reducer who is handling all the values for this particular key.

Every reducer input record then has a list of values and you can calculate (1+5+9)/3, (2+6+7)/3 and (3+4+8)/3. That’s all.

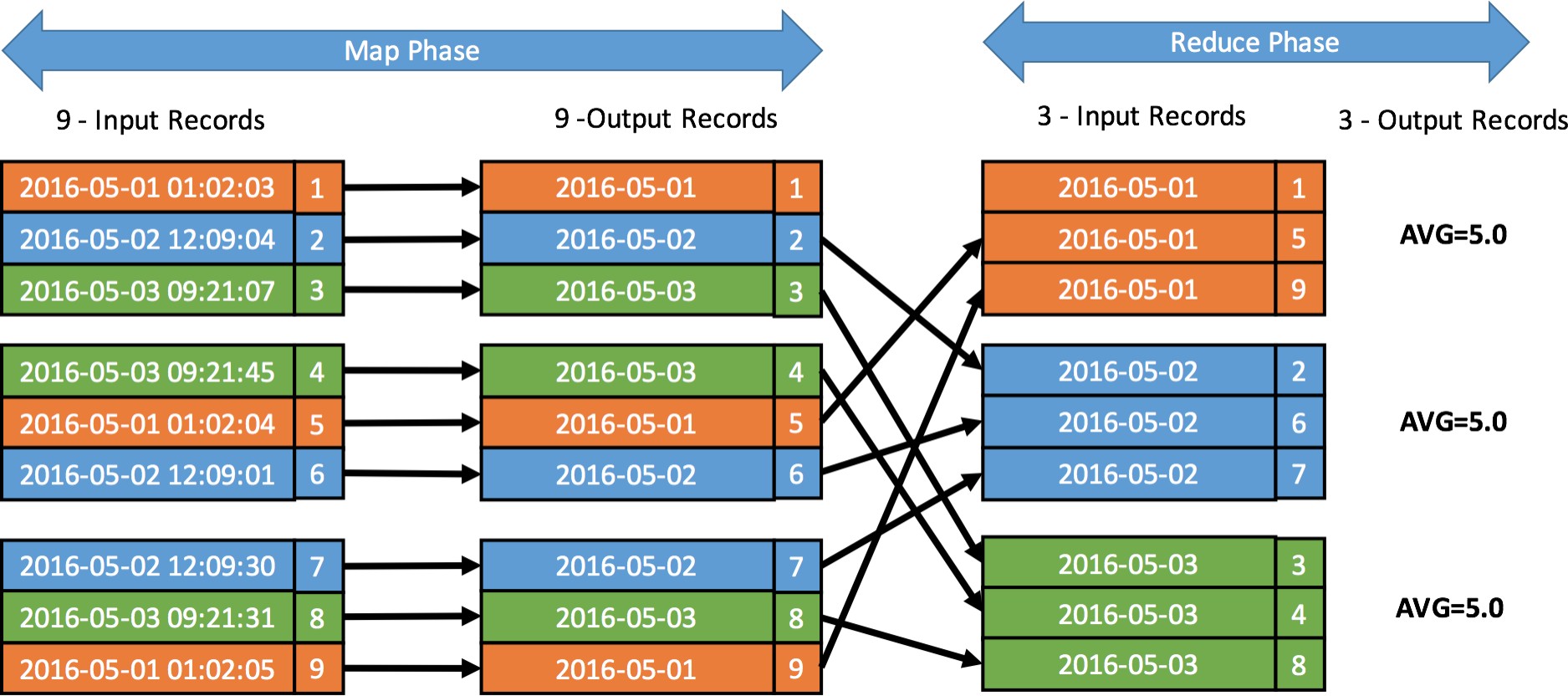


Figure 20.2: MapReduce Example of Time Series Data What do you think you need to do to generate minute averages?

Yes, you need to cut the key differently. You then would need to cut it like this: “2016- 05-01 01:02”. Keeping the Hour and minute information in the key.

What you can also see is, why map reduce is so great for doing parallel work. In this case, the map stage could be done by nine mappers in parallel because each map is independent from all the others.

The reduce stage could still be done by three tasks in parallel. One each for orange, blue and one for green.

That means, if your dataset would be 10 times as big and you’d have 10 times the machines, the time to do the calculation would be the same.

### What is the limitation of MapReduce? – available

MapReduce is awesome for simpler analytics tasks, like counting stuff. It just has one flaw: It has only two stages Map and Reduce.



Figure 20.3: The Map Reduce Process

First MapReduce loads the data from HDFS into the mapping function. There you prepare the input data for the processing in the reducer. After the reduce is finished the results get written to the data store.

The problem with MapReduce is that there is no simple way to chain multiple map and reduce processes together. At the end of each reduce process the data must be stored somewhere.

This fact makes it very hard to do complicated analytics processes. You would need to chain MapReduce jobs together.

Chaining jobs with storing and loading intermediate results just makes no sense.

Another issue with MapReduce is that it is not capable of streaming analytics. Jobs take some time to spin up, do the analytics and shut down. Basically Minutes of wait time are totally normal.

This is a big negative point in a more and more real time data processing world.

## Apache Spark

I talked about the three methods of data streaming in this podcast: https://anchor.fm/andreaskayy/emb Methods-of-Streaming-Data-e15r6o

### What is the difference to MapReduce? – available

Spark is a complete in memory framework. Data gets loaded from, for instance hdfs, into the memory of workers.

There is no longer a fixed map and reduce stage. Your code can be as complex as you want.

Once in memory, the input data and the intermediate results stay in memory (until the job finishes). They do not get written to a drive like with MapReduce.

This makes Spark the optimal choice for doing complex analytics. It allows you for instance to do Iterative processes. Modifying a dataset multiple times in order to create an output is totally easy.

Streaming analytics capability is also what makes spark so great. Spark has natively the option to schedule a job to run every X seconds or X milliseconds.

As a result, Spark can deliver you results from streaming data in “real time”.

### How does Spark fit to Hadoop? – available

There are some very misleading articles out there titled Spark or Hadoop, Spark is better than Hadoop or even Spark is replacing Hadoop.



Figure 20.4: Hadoop vs Spark capabilities

So, it’s time to show you the differences between Spark and Hadoop. After this you will know when and for what you should use Spark and Hadoop.

You’ll also understand why Hadoop or Spark is the totally wrong question.

### Where’s the difference?

To make it clear how Hadoop differs from Spark I created this simple feature table:

Hadoop is used to store data in the Hadoop Distributed File System (HDFS). It can analyse the stored data with MapReduce and manage resources with YARN.

However, Hadoop is more than just storage, analytics and resource management. There’s a whole eco system of tools around the Hadoop core. I’ve written about tis eco system in this article: What is Hadoop and why is it so freakishly popular. You should check it out as well.

Compared to Hadoop, Spark is “just” an analytics framework. It has no storage capa- bility. Although it has a standalone resource management, you usually don’t use that feature.

### Spark and Hadoop is a perfect fit

So, if Hadoop and Spark are not the same things, can they work together? Absolutely! Here’s how the first picture will look if you combine Hadoop with Spark:

As Storage you use the Hadoop distributed file system. Analytics is done with Apache Spark and Yarn is taking care of the resource management.

Why does that work so well together?

From a platform architecture perspective, Hadoop and Spark are usually managed on the same cluster. This means on each server where a HDFS data node is running, a spark worker thread runs as well.

In distributed processing, network transfer between machines is a large bottle neck. Transferring data within a machine reduces this traffic significantly.

Spark is able to determine on which data node the needed data is stored. This allows a direct load of the data from the local storage into the memory of the machine.

This reduces network traffic a lot.

### Spark on YARN:

You need to make sure that your physical resources are distributed perfectly between the services. This is especially the case when you run Spark workers with other Hadoop services on the same machine.

It just would not make sense to have two resource managers managing the same server’s resources. Sooner or later they will get in each others way.

That’s why the Spark standalone resource manager is seldom used.

So, the question is not Spark or Hadoop. The question has to be: Should you use Spark or MapReduce alongside Hadoop’s HDFS and YARN.

### My simple rule of thumb:

If you are doing simple batch jobs like counting values or doing calculating averages: Go with MapReduce.

If you need more complex analytics like machine learning or fast stream processing: Go with Apache Spark.

### Available Languages – available

Spark jobs can be programmed in a variety of languages. That makes creating analytic processes very user-friendly for data scientists.

Spark supports Python, Scala and Java. With the help of SparkR you can even connect your R program to a Spark cluster.

If you are a data scientist who is very familiar with Python just use Python, its great. If you know how to code Java I suggest you start using Scala.

Spark jobs are easier to code in Scala than in Java. In Scala you can use anonymous functions to do processing.

This results in less overhead, it is a much cleaner, simpler code.

With Java 8 simplified function calls were introduced with lambda expressions. Still, a lot of people, including me prefer Scala over Java.

### How to do stream processing

### How to do batch processing

### How does Spark use data from Hadoop – available

Another thing is data locality. I always make the point, that processing data locally where it is stored is the most efficient thing to do.

That’s exactly what Spark is doing. You can and should run Spark workers directly on the data nodes of your Hadoop cluster.

Spark can then natively identify on what data node the needed data is stored. This enables Spark to use the worker running on the machine where the data is stored to load

the data into the memory.

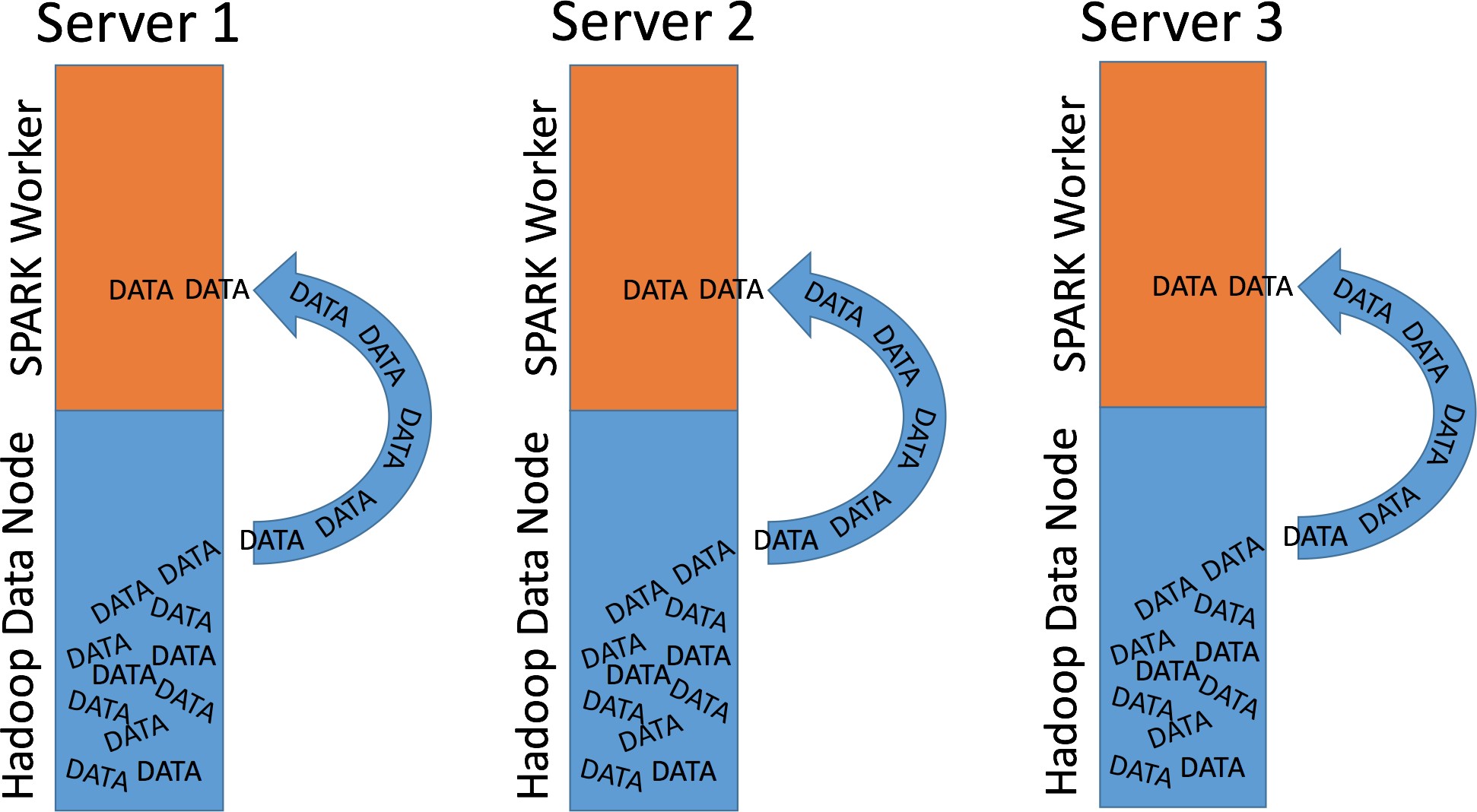


Figure 20.5: Spark Using Hadoop Data Locality

The downside of this setup is that you need more expensive servers. Because Spark processing needs stronger servers with more RAM and CPUs than a “pure” Hadoop setup.

### What is a RDD and what is a DataFrame?

### Spark coding with Scala

### Spark coding with Python

### How and why to use SparkSQL?

### Machine Learning on Spark? (Tensor Flow)

### MLlib:

The machine learning library MLlib is included in Spark so there is often no need to import another library.

I have to admit because I am not a data scientist I am not an expert in machine learning.

From what I have seen and read though the machine learning framework MLlib is a nice treat for data scientists wanting to train and apply models with Spark.

### Spark Setup – available

From a solution architect’s point of view Spark is a perfect fit for Hadoop big data platforms. This has a lot to do with cluster deployment and management.

Companies like Cloudera, MapR or Hortonworks include Spark into their Hadoop distri- butions. Because of that, Spark can be deployed and managed with the clusters Hadoop management web fronted.

This makes the process for deploying and configuring a Spark cluster very quick and admin friendly.

### Spark Resource Management – available

When running a computing framework you need resources to do computation: CPU time, RAM, I/O and so on. Out of the box Spark can manage resources with it’s stand-alone resource manager.

If Spark is running in an Hadoop environment you don’t have to use Spark’s own stand- alone resource manager. You can configure Spark to use Hadoop’s YARN resource man- agement.

Why would you do that? It allows YARN to efficiently allocate resources to your Hadoop and Spark processes.

Having a single resource manager instead of two independent ones makes it a lot easier to configure the resource management.

## Apache Nifi

Great blog about Nifi: [https://www.datainmotion.dev](https://www.datainmotion.dev/)

## StreamSets

<https://youtu.be/djt8532UWow>

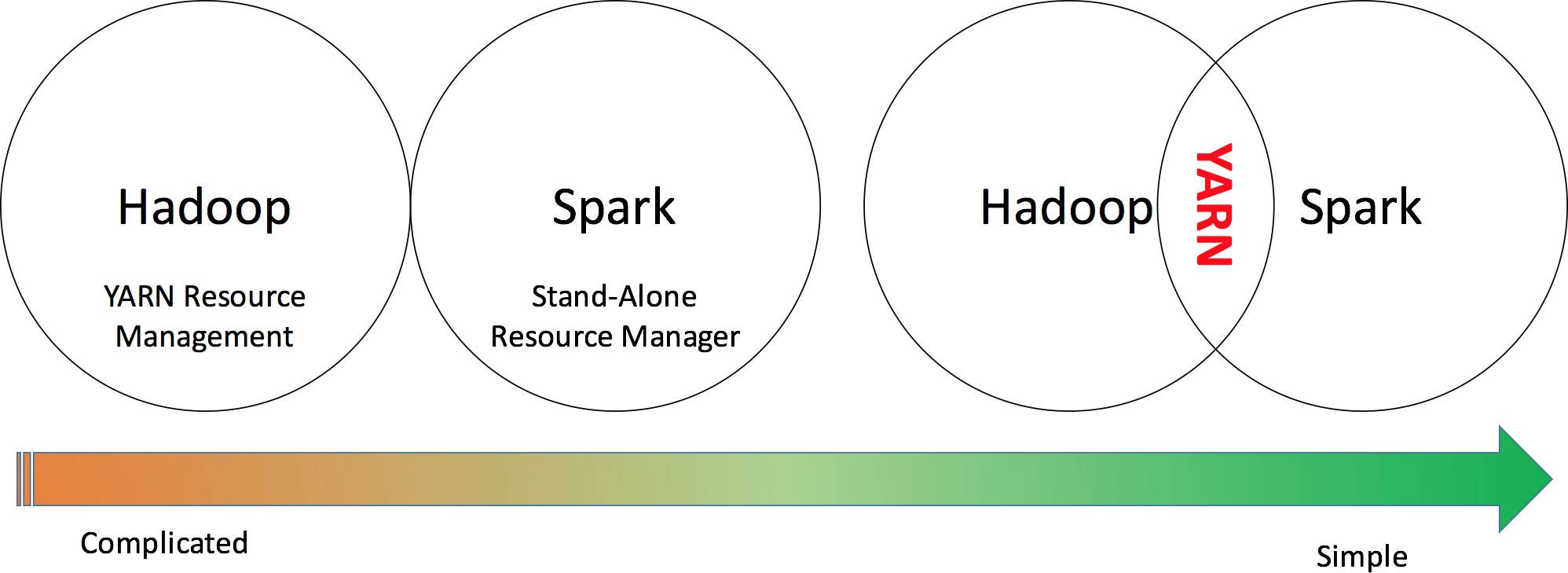


Figure 20.6: Spark Resource Management With YARN

[https://www.youtube.com/watch?v=Qm5e574WoCU&t=2s](https://www.youtube.com/watch?v=Qm5e574WoCU&amp;t=2s)

[https://github.com/gschmutz/stream-processing-workshop/tree/master/04-twitter-data-ingestion-wit](https://github.com/gschmutz/stream-processing-workshop/tree/master/04-twitter-data-ingestion-with-streamsets) <https://streamsets.com/blog/streaming-data-twitter-analysis-spark/>

# Apache Kafka

## Why a message queue tool?

## Kakfa architecture

## What are topics

## What does Zookeeper have to do with Kafka

## How to produce and consume messages

My YouTube video how to set up Kafka at home: https://youtu.be/7F9tBwTUSeY My YouTube video how to write to Kafka: https://youtu.be/RboQBZvZCh0

## KAFKA Commands

Start Zookeeper container for Kafka:

*\*

*\*

docker run *−*d ***−*** name zookeeper *−*s e rv e r

***−*** network app*−*t i e r *\*

*−*e ALLOW ANONYMOUS LOGIN=y es

bitnami / zookeeper : l a t e s t

Start Kafka container:

docker run *−*d ***−*** name kafka*−*s e rv e r

***−*** network app*−*t i e r *\*

*\*

*−*e KAFKA CFG ZOOKEEPER CONNECT=z ookeeper *−*s e rv e r : 2181

*\*

*−*e ALLOW PLAINTEXT LISTENER=y es *\*

bitnami / kafka : l a t e s t

# Machine Learning

**Podcast Episode:** Machine Learning In Production

Doing machine learning in production is very different then for proof of concepts or in education. One of the hardest parts is keeping models updated.

Anchor [Click here to listen](https://anchor.fm/andreaskayy/episodes/Machine-Learning-In-Production-e11bbk)

Table 22.1: Podcast: Machine Learning In Production

## Training and Applying models

## What is deep learning

## How to do Machine Learning in production — available

Machine learning in production is using stream and batch processing. In the batch processing layer you are creating the models, because you have all the data available for training.

In the stream in processing layer you are using the created models, you are applying them to new data.

The idea that you need to incorporate is that it is a constant cycle. Training, applying, re-training, pushing into production and applying.

What you don’t want to do is, you don’t want to do this manually. You need to figure out a process of automatic retraining and automatic pushing to into production of models.

In the retraining phase the system automatically evaluates the training. If the model no longer fits it works as long as it needs to create a good model.

After the evaluation of the model is complete and it’s good, the model gets pushed into production. Into the stream processing.

## Why machine learning in production is harder then you think – available

How to automate machine learning is something that drives me day in and day out.

What you do in development or education is, that you create a model and fit it to the data. Then that model is basically done forever.

Where I’m coming from, the IoT world, the problem is that machines are very different. They behave very different and experience wear.

## Models Do Not Work Forever

Machines have certain processes that decrease the actual health of the machine. Machine wear is a huge issue. Models that that are built on top of a good machine don’t work forever.

When the Machine wears out, the models need to be adjusted. They need to be main- tained, retrained.

## Where The Platforms That Support This?

Automatic re-training and re-deploying is a very big issue, a very big problem for a lot of companies. Because most existing platforms don’t have this capability (I actually haven’t seen one until now).

Look at AWS machine learning for instance. The process is: build, train, tune deploy. Where’s the loop of retraining?

You can create models and then use them in production. But this loop is almost nowhere to be seen.

It is a very big issue that needs to be solved. If you want to do machine learning in production you can start with manual interaction of the training, but at some point you need to automate everything.

## Training Parameter Management

To train a model you are manipulating input parameters of the models. Take deep learning for instance. To train you are manipulating for instance:

How many layers do you use. The depth of the layers, which means how many neurons you have in a layer. What activation function you use, how long are you training and so on.

You also need to keep track of what data you used to train which model.

All those parameters need to be manipulated automatically, models trained and tested. To do all that, you basically need a database that keeps track of those variables.

How to automate this, for me, is like the big secret. I am still working on figuring it out.

## What’s Your Solution?

Did you already have the problem of automatic re-training and deploying of models as well?

Were you able to use a cloud platform like Google, AWS or Azure? It would be really awesome if you share your experience :)

## How to convince people machine learning works

**— available**

Many people still are not convinced that machine learning works reliably. But they want analytics insight and most of the time machine learning is the way to go.

This means, when you are working with customers you need to do a lot of convincing. Especially if they are not into machine learning themselves.

But it’s actually quite easy.

## No Rules, No Physical Models

Many people are still under the impression that analytics only works when it’s based on physics. When there are strict mathematical rules to a problem.

Especially in engineering heavy countries like Germany this is the norm:

“Sere has to be a Rule for Everysing!” (imagine a German accent) When you’re engi- neering you are calculating stuff based on physics and not based on data. If you are constructing an airplane wing, you better make sure to use calculations so it doesn’t fall off.

And that’s totally fine.

Keep doing that!

Machine learning has been around for decades. It didn’t quite work as good as people hoped. We have to admit that. But there is this preconception that it still doesn’t work.

Which is not true: Machine learning works.

Somehow you need to convince people that it is a viable approach. That learning from data to make predictions is working perfectly.

## You Have The Data. USE IT!

As a data scientist you have one ace up your sleeve, it’s the obvious one: It’s the data and it’s statistics.

You can use that data and those statistics to counter peoples preconceptions. It’s very powerful if someone says: “This doesn’t work”

You bring the data. You show the statistics and you show that it works reliably. A lot of discussions end there.

Data doesn’t lie. You can’t fight data. The data is always right.

## Data is Stronger Than Opinions

This is also why I believe that autonomous driving will come quicker than many of us think. Because a lot of people say, they are not safe. That you cannot rely on those cars.

The thing is: When you have the data you can do the statistics.

You can show people that autonomous driving really works reliably. You will see, the question of: Is this is this allowed or is this not allowed? Will be gone quicker than you think.

Because government agencies can start testing the algorithms based on predefined sce- narios. They can run benchmarks and score the cars performance.

All those opinions, if it works, or if it doesn’t work, they will be gone.

The motor agency has the statistics. The stats show people how good cars work. Companies like Tesla, they have it very easy. Because the data is already there. **They just need to show us that the algorithms work. The end.**

## AWS Sagemaker

Train and apply models online with Sagemaker

Link to the OLX Slideshare with pros, cons and how to use Sagemaker: [https://www.](https://www.slideshare.net/mobile/AlexeyGrigorev/image-models-infrastructure-at-olx) [slideshare.net/mobile/AlexeyGrigorev/image-models-infrastructure-at-olx](https://www.slideshare.net/mobile/AlexeyGrigorev/image-models-infrastructure-at-olx)

# Data Visualization

## Android & IOS

## How to design APIs for mobile apps

## How to use Webservers to display content

This section does not contain any text that’s why the page is messed up

### Tomcat

* + 1. **Jetty**
    2. **NodeRED**
    3. **React**

## Business Intelligence Tools

* + 1. **Tableau**
    2. **PowerBI**
    3. **Quliksense**

## Identity & Device Management

* + 1. **What is a digital twin?**
    2. **Active Directory**

**Part III**

**Data Engineering Course: Building A Data Platform**

# What We Want To Do

* Twitter data to predict best time to post using the hashtag datascience or ai
* Find top tweets for the day
* Top users
* Analyze sentiment and keywords

# Thoughts On Choosing A Development Environment

For a local environment you need a good PC. I thought a bit about a budget build around

1.000 Dollars or Euros.

**Podcast Episode:** #068 How to Build a Budget Data Science PC

In this podcast we look into configuring a sub 1000 dollar PC for data engineering and machine learning

YouTube [Click here to watch](https://youtu.be/00NWR-II6ek) Audio [Click here to listen](https://anchor.fm/andreaskayy/episodes/068-A-Budget-Data-Science-PC-Build-e45inh)

Table 25.1: Podcast: 068 How to Build a Budget Data Science PC

# A Look Into the Twitter API

**Podcast Episode:** #081 Twitter API Research

In this podcast we were looking into how the Twitter API works and how you get access to it

YouTube [Click here to watch](https://youtu.be/UnAXKxeIlyg)

Table 26.1: Podcast: 081 Twitter API Research

# Ingesting Tweets with Apache Nifi

**Podcast Episode:** #082 Reading Tweets With Apache Nifi & IaaS vs PaaS vs SaaS

In this podcast we are trying to read Twitter Data with Nifi

YouTube [Click here to watch](https://youtu.be/pWuT4UAocUY)

Table 27.1: Podcast: 082 Reading Tweets With Apache Nifi

**Podcast Episode:** #085 Trying to read Tweets with Nifi Part 2

We are looking into the Big Data landscape chart and we are trying to read Twitter Data with Nifi again

YouTube [Click here to watch](https://youtu.be/OLUwXr8-gAk)

Table 27.2: Podcast: 085 Trying to read Tweets with Nifi Part 2

# Writing from Nifi to Apache Kafka

**Podcast Episode:** #086 How to Write from Nifi to Kafka Part 1

I’ve been working a lot on the cookbook, because it’s so much fun. I gotta tell you what I added. Then we are trying to write the Tweets from Apache Nifi into Kafka. Also talk about Kafka basics.

YouTube [Click here to watch](https://youtu.be/F7Y-ygnyJMg)

Table 28.1: Podcast: 086 How to Write from Nifi to Kafka Part 1

**Podcast Episode:** #088 How to Write from Nifi to Kafka Part 2

In this podcast we finally figure out how to write to Kafka from Nifi. The problem was the network configuration of the Docker containers

YouTube [Click here to watch](https://youtu.be/pJbRnBQmoCs)

Table 28.2: Podcast: 088 How to Write from Nifi to Kafka Part 2

# Apache Zeppelin

## Install and Ingest Kafka Topic

Start the container:

docker run *−*d *−*p 8081 : 8080 ***−*** rm *\*

*−*v / Users / xxxx/Documents/ Docker Files / l o g s : / l o g s *\*

*−*v / Users / xxxx/Documents/ Docker Files / Notebooks : / notebook *\*

*−*e ZEPPELIN LOG DIR=’/ logs ’ *\*

*−*e ZEPPELIN NOTEBOOK DIR=’/ notebook ’ *\*

***−*** network app*−*t i e r ***−*** name z e ppe l i n apache / z e ppe l i n : 0 . 7 . 3

## Processing Messages with Spark & SparkSQL

## Visualizing Data

# Switch Processing from Zeppelin to Spark

## Install Spark

## Ingest Messages from Kafka

## Writing from Spark to Kafka

## Move Zeppelin Code to Spark

**Part IV Case Studies**

# How I do Case Studies

## Data Science @Airbnb

**Podcast Episode:** #063 Data Engineering At Airbnb Case Study

How Airbnb is doing Data Engineering? Let’s check it out.

YouTube [Click here to watch](https://youtu.be/iokqkMfyIfo) Audio [Click here to listen](https://anchor.fm/andreaskayy/episodes/063-Data-Engineering-At-Airbnb-Case-Study-e45il2)

Table 31.1: Podcast: 063 Data Engineering At Airbnb Case Study

#### Slides:

<https://medium.com/airbnb-engineering/airbnb-engineering-infrastructure/home> Airbnb Engineering Blog: <https://medium.com/airbnb-engineering>

Data Infrastructure: [https://medium.com/airbnb-engineering/data-infrastructure-at-airbnb-8adfb34f1](https://medium.com/airbnb-engineering/data-infrastructure-at-airbnb-8adfb34f169c)

Scaling the serving tier: [https://medium.com/airbnb-engineering/unlocking-horizontal-scalability-in-ou](https://medium.com/airbnb-engineering/unlocking-horizontal-scalability-in-our-web-serving-tier-d907449cdbcf)

Druid Analytics: <https://medium.com/airbnb-engineering/druid-airbnb-data-platform-601c312f2a4c>

Spark Streaming for logging events: [https://medium.com/airbnb-engineering/scaling-spark-streaming-f](https://medium.com/airbnb-engineering/scaling-spark-streaming-for-logging-event-ingestion-4a03141d135d)

-Druid Wiki: [https://en.wikipedia.org/wiki/Apache Druid](https://en.wikipedia.org/wiki/Apache_Druid)

**31.1.1 Data Science @Amazon**

[https://www.datasciencecentral.com/profiles/blogs/20-data-science-systems-used-by-amazon-to-opera](https://www.datasciencecentral.com/profiles/blogs/20-data-science-systems-used-by-amazon-to-operate-its-business) <https://aws.amazon.com/solutions/case-studies/amazon-migration-analytics/>

## Data Science @Baidu

[https://www.slideshare.net/databricks/spark-sql-adaptive-execution-unleashes-the-power-of-cluster-in](https://www.slideshare.net/databricks/spark-sql-adaptive-execution-unleashes-the-power-of-cluster-in-large-scale-with-chenzhao-guo-and-carson-wang)

## Data Science @Blackrock

[https://www.slideshare.net/DataStax/maintaining-consistency-across-data-centers-randy-fradin-black](https://www.slideshare.net/DataStax/maintaining-consistency-across-data-centers-randy-fradin-blackrock-cassandra-summit-2016)

## Data Science @BMW

Big Data in der Automobilindustrie – Daten aus dem Fahrzeug nutzen [https://www.](https://www.unibw.de/code.../ws3_bigdata_vortrag_widmann.pdf) [unibw.de/code.../ws3 bigdata vortrag widmann.pdf](https://www.unibw.de/code.../ws3_bigdata_vortrag_widmann.pdf)

## Data Science @Booking.com

**Podcast Episode:** #064 Data Engineering At Booking.com Case Study

How Booking.com is doing Data Engineering? Let’s check it out.

YouTube [Click here to watch](https://youtu.be/9GE3yiVo1FM) Audio [Click here to listen](https://anchor.fm/andreaskayy/episodes/064-Data-Engineering-At-Booking-com-Case-Study-e45ilg)

Table 31.2: Podcast: 064 Data Engineering At Booking.com Case Study

#### Slides:

h[ttps://www.slideshare.net/ConfluentInc/data-](http://www.slideshare.net/ConfluentInc/data-streaming-ecosystem-management-at-bookingcom)st[reaming-ecosystem-](http://www.slideshare.net/ConfluentInc/data-streaming-ecosystem-management-at-bookingcom)m[anagemen](http://www.slideshare.net/ConfluentInc/data-streaming-ecosystem-management-at-bookingcom)t-[at-](http://www.slideshare.net/ConfluentInc/data-streaming-ecosystem-management-at-bookingcom)bo[okingcom?](http://www.slideshare.net/ConfluentInc/data-streaming-ecosystem-management-at-bookingcom) ref=h[ttps://www](http://www.confluent.io/kafka-summit-sf18/data-streaming-ecosystem-management).confl[uent.io/kafka-summit-sf18/data-streaming-ecosystem-managemen](http://www.confluent.io/kafka-summit-sf18/data-streaming-ecosystem-management)t

[https://www.slideshare.net/SparkSummit/productionizing-behavioural-features-for-machine-learning-](https://www.slideshare.net/SparkSummit/productionizing-behavioural-features-for-machine-learning-with-apache-spark-streaming-with-ben-teeuwen-and-roman-studenikin)

h[ttps://www.slideshare.net/ConfluentInc/data-](http://www.slideshare.net/ConfluentInc/data-streaming-ecosystem-management-at-bookingcom)st[reaming-ecosystem-](http://www.slideshare.net/ConfluentInc/data-streaming-ecosystem-management-at-bookingcom)m[anagemen](http://www.slideshare.net/ConfluentInc/data-streaming-ecosystem-management-at-bookingcom)t-[at-](http://www.slideshare.net/ConfluentInc/data-streaming-ecosystem-management-at-bookingcom)bo[okingcom?](http://www.slideshare.net/ConfluentInc/data-streaming-ecosystem-management-at-bookingcom) ref=h[ttps://www](http://www.confluent.io/kafka-summit-sf18/data-streaming-ecosystem-management).confl[uent.io/kafka-summit-sf18/data-streaming-ecosystem-managemen](http://www.confluent.io/kafka-summit-sf18/data-streaming-ecosystem-management)t

Druid: <https://towardsdatascience.com/introduction-to-druid-4bf285b92b5a> Kafka Architecture: <https://data-flair.training/blogs/kafka-architecture/> Confluent Platform: <https://www.confluent.io/product/confluent-platform/>

## Data Science @CERN

#### Slides:

[https://en.wikipedia.org/wiki/Large Hadron Collider](https://en.wikipedia.org/wiki/Large_Hadron_Collider) <http://www.lhc-facts.ch/index.php?page=datenverarbeitung>

**Podcast Episode:** #065 Data Engineering At CERN Case Study

How is CERN doing Data Engineering? The must get huge amounts of data from the Large Hydron Colider. Let’s check it out.

YouTube [Click here to watch](https://youtu.be/LrhfzPsKaDE) Audio [Click here to listen](https://anchor.fm/andreaskayy/episodes/065-Data-Engineering-At-CERN-Case-Study-e45ime)

Table 31.3: Podcast: 065 Data Engineering At CERN Case Study

[https://openlab.cern/sites/openlab.web.cern.ch/files/2018-09/2017 ESADE Madrid Big](https://openlab.cern/sites/openlab.web.cern.ch/files/2018-09/2017_ESADE_Madrid_Big_Data.pdf) [Data.pdf](https://openlab.cern/sites/openlab.web.cern.ch/files/2018-09/2017_ESADE_Madrid_Big_Data.pdf)

[https://openlab.cern/sites/openlab.web.cern.ch/files/2018-05/kubeconeurope2018-cern-](https://openlab.cern/sites/openlab.web.cern.ch/files/2018-05/kubeconeurope2018-cern-180507122303.pdf)180507122303. [pdf](https://openlab.cern/sites/openlab.web.cern.ch/files/2018-05/kubeconeurope2018-cern-180507122303.pdf)

<https://www.slideshare.net/SparkSummit/next-cern-accelerator-logging-service-with-jakub-wozniak> <https://databricks.com/session/the-architecture-of-the-next-cern-accelerator-logging-service> [http://opendata.cern.ch](http://opendata.cern.ch/)

[https://gobblin.apache.org](https://gobblin.apache.org/)

[https://www.slideshare.net/databricks/cerns-next-generation-data-analysis-platform-with-apache-spa](https://www.slideshare.net/databricks/cerns-next-generation-data-analysis-platform-with-apache-spark-with-enric-tejedor)

[https://www.slideshare.net/SparkSummit/realtime-detection-of-anomalies-in-the-database-infrastruct](https://www.slideshare.net/SparkSummit/realtime-detection-of-anomalies-in-the-database-infrastructure-using-apache-spark-with-daniel-lanza-and-prasanth-kothuri)

## Data Science @Disney

[https://medium.com/disney-streaming/delivering-data-in-real-time-via-auto-scaling-kinesis-streams-7](https://medium.com/disney-streaming/delivering-data-in-real-time-via-auto-scaling-kinesis-streams-72a0236b2cd9)

## Data Science @Drivetribe

[https://berlin-2017.flink-forward.org/kb sessions/drivetribes-kappa-architecture-with-apache-flink/](https://berlin-2017.flink-forward.org/kb_sessions/drivetribes-kappa-architecture-with-apache-flink/) [https://www.slideshare.net/FlinkForward/flink-forward-berlin-2017-aris-kyriakos-koliopoulos-drivetri](https://www.slideshare.net/FlinkForward/flink-forward-berlin-2017-aris-kyriakos-koliopoulos-drivetribes-kappa-architecture-with-apache-flink)

## Data Science @Dropbox

<https://blogs.dropbox.com/tech/2019/01/finding-kafkas-throughput-limit-in-dropbox-infrastructure/>

## Data Science @Ebay

[https://www.slideshare.net/databricks/moving-ebays-data-warehouse-over-to-apache-spark-spark-as-c](https://www.slideshare.net/databricks/moving-ebays-data-warehouse-over-to-apache-spark-spark-as-core-etl-platform-at-ebay-with-kim-curtis-and-brian-knauss) [https://www.slideshare.net/databricks/analytical-dbms-to-apache-spark-auto-migration-framework-w](https://www.slideshare.net/databricks/analytical-dbms-to-apache-spark-auto-migration-framework-with-edward-zhang-and-lipeng-zhu)

## Data Science @Expedia

<https://www.slideshare.net/BrandonOBrien/spark-streaming-kafka-best-practices-w-brandon-obrien> <https://www.slideshare.net/Naveen1914/brandon-obrien-streamingdata>

## Data Science @Facebook

<https://code.fb.com/core-data/apache-spark-scale-a-60-tb-production-use-case/>

## Data Science @Google

<http://www.unofficialgoogledatascience.com/> <https://ai.google/research/teams/ai-fundamentals-applications/> <https://cloud.google.com/solutions/big-data/> <https://datafloq.com/read/google-applies-big-data-infographic/385>

## Data Science @@Grammarly

[https://www.slideshare.net/databricks/building-a-versatile-analytics-pipeline-on-top-of-apache-spark-](https://www.slideshare.net/databricks/building-a-versatile-analytics-pipeline-on-top-of-apache-spark-with-mikhail-chernetsov)

## Data Science @ING Fraud

[https://sf-2017.flink-forward.org/kb sessions/streaming-models-how-ing-adds-models-at-runtime-to-ca](https://sf-2017.flink-forward.org/kb_sessions/streaming-models-how-ing-adds-models-at-runtime-to-catch-fraudsters/)

## Data Science @Instagram

[https://www.slideshare.net/SparkSummit/lessons-learned-developing-and-managing-massive-300tb-ap](https://www.slideshare.net/SparkSummit/lessons-learned-developing-and-managing-massive-300tb-apache-spark-pipelines-in-production-with-brandon-carl)

## Data Science @LinkedIn

**Podcast Episode:** #073 Data Engineering At LinkedIn Case Study

Let’s check out how LinkedIn is processing data :)

YouTube [Click here to watch](https://youtu.be/wgfoE8Jbr_Q) Audio [Click here to listen](https://anchor.fm/andreaskayy/episodes/073-Data-Engineering-At-LinkedIn-Case-Study-e45is6)

Table 31.4: Podcast: 073 Data Engineering At LinkedIn Case Study

#### Slides:

<https://engineering.linkedin.com/teams/data#0>

[https://www.slideshare.net/yaelgarten/building-a-healthy-data-ecosystem-around-kafka-and-hadoop-l](https://www.slideshare.net/yaelgarten/building-a-healthy-data-ecosystem-around-kafka-and-hadoop-lessons-learned-at-linkedin)

<https://engineering.linkedin.com/teams/data/projects/pinot> [https://pinot.readthedocs.io/en/latest/intro.html#](https://pinot.readthedocs.io/en/latest/intro.html) <https://towardsdatascience.com/building-machine-learning-at-linkedin-scale-f08bd9a63f0a> [http://samza.apache.org](http://samza.apache.org/)

h[ttps://www.slideshare.net/ConfluentInc/more-](http://www.slideshare.net/ConfluentInc/more-data-more-problems-scaling-kafkamirroring-pipelines-)data-[more-problems-scaling-](http://www.slideshare.net/ConfluentInc/more-data-more-problems-scaling-kafkamirroring-pipelines-)k[afkamirroring-pipelines-](http://www.slideshare.net/ConfluentInc/more-data-more-problems-scaling-kafkamirroring-pipelines-) ref=h[ttps://www](http://www.confluent.io/kafka-summit-sf18/more).confl[uent.io/kafka-summit-sf18/more](http://www.confluent.io/kafka-summit-sf18/more) data more problems

[https://www.slideshare.net/KhaiTran17/conquering-the-lambda-architecture-in-linkedin-metrics-platf](https://www.slideshare.net/KhaiTran17/conquering-the-lambda-architecture-in-linkedin-metrics-platform-with-apache-calcite-and-apache-samza)

[https://www.slideshare.net/Hadoop Summit/unified-batch-stream-processing-with-apache-samza](https://www.slideshare.net/Hadoop_Summit/unified-batch-stream-processing-with-apache-samza) <http://druid.io/docs/latest/design/index.html>

## Data Science @Lyft

<https://eng.lyft.com/running-apache-airflow-at-lyft-6e53bb8fccff>

## Data Science @NASA

#### Slides:

[http://sites.nationalacademies.org/cs/groups/ssbsite/documents/webpage/ssb 182893.pdf](http://sites.nationalacademies.org/cs/groups/ssbsite/documents/webpage/ssb_182893.pdf) [https://esip.figshare.com/articles/Apache Science Data Analytics Platform/5786421](https://esip.figshare.com/articles/Apache_Science_Data_Analytics_Platform/5786421)

**Podcast Episode:** #067 Data Engineering At NASA Case Study A look into how NASA is doing data engineering.

YouTube [Click here to watch](https://youtu.be/Pctn_1UoNjA) Audio [Click here to listen](https://anchor.fm/andreaskayy/episodes/067-Data-Engineering-At-NASA-Case-Study-e45ina)

Table 31.5: Podcast: 067 Data Engineering At NASA Case Study

[http://www.socallinuxexpo.org/sites/default/files/presentations/OnSightCloudArc](http://www.socallinuxexpo.org/sites/default/files/presentations/OnSightCloudArchitecture-scale14x.pdf)hitecture-scale14x. [pdf](http://www.socallinuxexpo.org/sites/default/files/presentations/OnSightCloudArchitecture-scale14x.pdf)

[https://www.slideshare.net/SparkSummit/spark-at-nasajplchris-mattmann?qid=909685](https://www.slideshare.net/SparkSummit/spark-at-nasajplchris-mattmann?qid=90968554-288e-454a-b63a-21a45cfc897d&amp;v&amp;b&amp;from_search=4)54-288e-454a- [v=&b=&from search=4](https://www.slideshare.net/SparkSummit/spark-at-nasajplchris-mattmann?qid=90968554-288e-454a-b63a-21a45cfc897d&amp;v&amp;b&amp;from_search=4)

[https://en.m.wikipedia.org/wiki/Hierarchical Data Format](https://en.m.wikipedia.org/wiki/Hierarchical_Data_Format)

## Data Science @Netflix – available

**Podcast Episode:** #062 Data Engineering At Netflix Case Study

How Netflix is doing Data Engineering using their Keystone platform.

YouTube [Click here to watch](https://youtu.be/YWPsYpjNKeM) Audio [Click here to listen](https://anchor.fm/andreaskayy/episodes/062-Data-Engineering-At-Netflix-Case-Study-e45ikp)

Table 31.6: Podcast: 062 Data Engineering At Netflix Case Study

#### Slides:

Netflix revolutionized how we watch movies and tv. Currently over 75 million users watch 125 million hours of Netflix content every day!

Netflix’s revenue comes from a monthly subscription service. So, the goal for Netflix is to keep you subscribed and to get new subscribers.

To achieve this, Netflix is licensing movies from studios as well as creating its own original movies and tv series.

But offering new content is not everything. What is also very important is, to keep you watching content that already exists.

To be able to recommend you content, Netflix is collecting data from users. And it is collecting a lot.

Currently, Netflix analyses about 500 billion user events per day. That results in a stunning 1.3 petabytes every day.

All this data allows Netflix to build recommender systems for you. The recommenders are showing you content that you might like, based on your viewing habits, or what is currently trending.

**The Netflix batch processing pipeline** When Netflix started out, they had a very simple batch processing system architecture.

The key components were Chuckwa, a scalable data collection system, Amazon S3 and Elastic MapReduce.

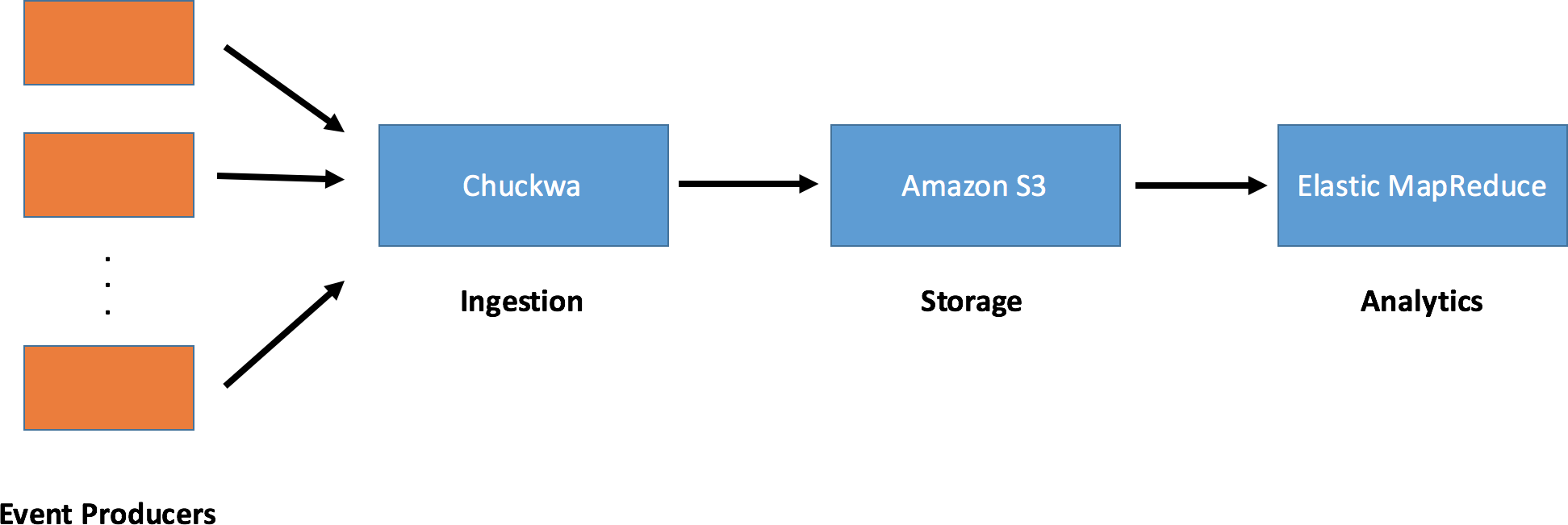


Figure 31.1: Old Netflix Batch Processing Pipeline

Chuckwa wrote incoming messages into Hadoop sequence files, stored in Amazon S3. These files then could be analysed by Elastic MapReduce jobs.

Netflix batch processing pipeline Jobs were executed regularly on a daily and hourly basis. As a result, Netflix could learn how people used the services every hour or once a day.

**Know what customers want:** Because you are looking at the big picture you can create new products. Netflix uses insight from big data to create new tv shows and movies.

They created House of Cards based on data. There is a very interesting Ted talk about this you should watch:

How to use data to make a hit TV show — Sebastian Wernicke:

Batch processing also helps Netflix to know the exact episode of a TV show that gets you hooked. Not only globally but for every country where Netflix is available.

Check out the article from TheVerge

They know exactly what show works in what country and what show does not.

It helps them create shows that work in everywhere or select the shows to license in different countries. Germany for instance does not have the full library that Americans have :(

We have to put up with only a small portion of tv shows and movies. If you have to select, why not select those that work best.

**Batch processing is not enough** As a data platform for generating insight the Cuckwa pipeline was a good start. It is very important to be able to create hourly and daily aggregated views for user behavior.

To this day Netflix is still doing a lot of batch processing jobs.

The only problem is: With batch processing you are basically looking into the past.

For Netflix, and data driven companies in general, looking into the past is not enough. They want a live view of what is happening.

**The trending now feature** One of the newer Netflix features is “Trending now”. To the average user it looks like that “Trending Now” means currently most watched.

This is what I get displayed as trending while I am writing this on a Saturday morning at 8:00 in Germany. But it is so much more.

What is currently being watched is only a part of the data that is used to generate “Trending Now”.



Figure 31.2: Netflix Trending Now Feature

“Trending now” is created based on two types of data sources: Play events and Impression events.

What messages those two types actually include is not really communicated by Netflix. I did some research on the Netflix Techblog and this is what I found out:

Play events include what title you have watched last, where you did stop watching, where you used the 30s rewind and others. Impression events are collected as you browse the

Netflix Library like scroll up and down, scroll left or right, click on a movie and so on

Basically, play events log what you do while you are watching. Impression events are capturing what you do on Netflix, while you are not watching something.

**Netflix real-time streaming architecture** Netflix uses three internet facing services to exchange data with the client’s browser or mobile app. These services are simple Apache Tomcat based web services.

The service for receiving play events is called “Viewing History”. Impression events are collected with the “Beacon” service.

The “Recommender Service” makes recommendations based on trend data available for clients.

Messages from the Beacon and Viewing History services are put into Apache Kafka. It acts as a buffer between the data services and the analytics.

Beacon and Viewing History publish messages to Kafka topics. The analytics subscribes to the topics and gets the messages automatically delivered in a first in first out fashion.

After the analytics the workflow is straight forward. The trending data is stored in a Cassandra Key-Value store. The recommender service has access to Cassandra and is making the data available to the Netflix client.

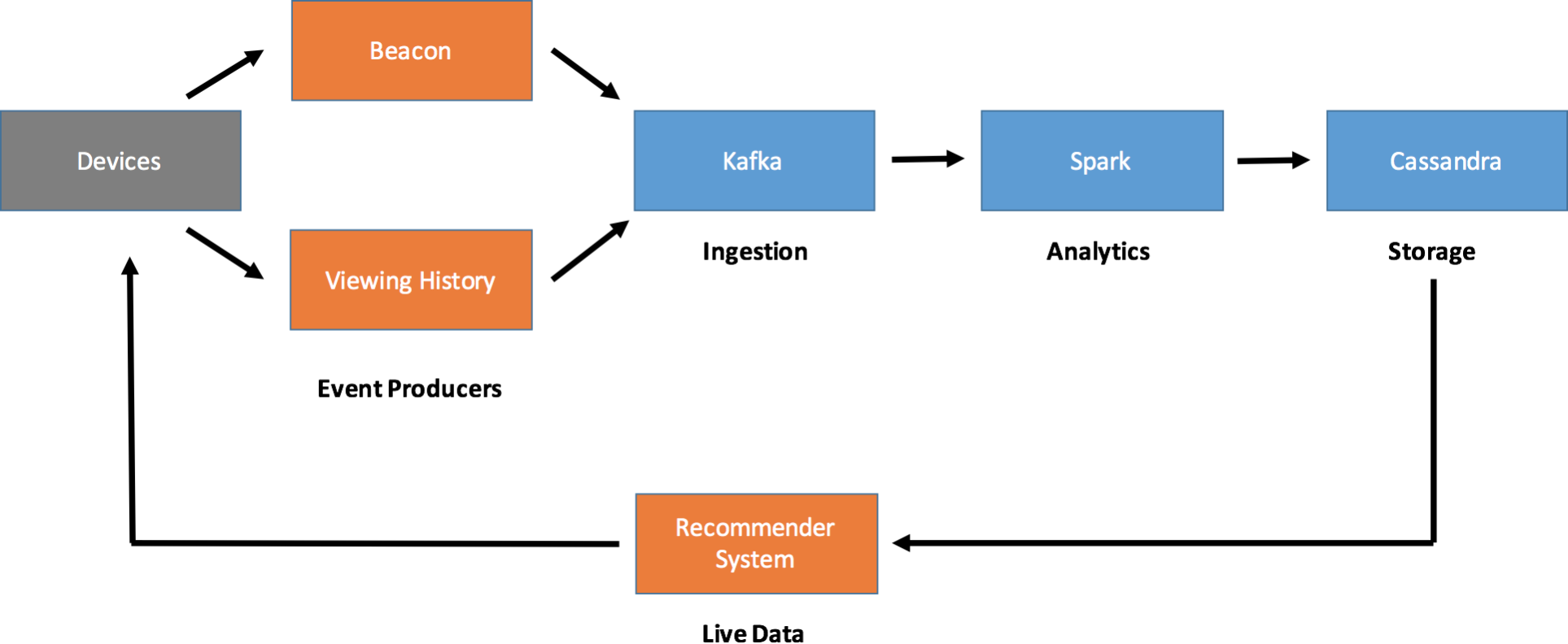


Figure 31.3: Netflix Streaming Pipeline

The algorithms how the analytics system is processing all this data is not known to the public. It is a trade secret of Netflix.

What is known, is the analytics tool they use. Back in Feb. 2015 they wrote in the tech blog that they use a custom made tool.

They also stated, that Netflix is going to replace the custom made analytics tool with Apache Spark streaming in the future. My guess is, that they did the switch to Spark some time ago, because their post is more than a year old.

## Data Science @OLX

**Podcast Episode:** #083 Data Engineering at OLX Case Study

This podcast is a case study about OLX with Senior Data Scientist Alexey Grigorev as guest. It was super fun.

YouTube [Click here to watch](https://youtu.be/H_uFNoCvykM) Audio [Click here to listen](https://anchor.fm/andreaskayy/episodes/083-Data-Engineering-at-OLX-Case-Study-e45j5n)

Table 31.7: Podcast: 083 Data Engineering at OLX Case Study

#### Slides:

<https://www.slideshare.net/mobile/AlexeyGrigorev/image-models-infrastructure-at-olx>

## Data Science @OTTO

[https://www.slideshare.net/SparkSummit/spark-summit-eu-talk-by-sebastian-schroeder-and-ralf-sigm](https://www.slideshare.net/SparkSummit/spark-summit-eu-talk-by-sebastian-schroeder-and-ralf-sigmund)

## Data Science @Paypal

<https://www.paypal-engineering.com/tag/data/>

## Data Science @Pinterest

**Podcast Episode:** #069 Engineering Culture At Pinterest

In this podcast we look into data platform and processing at Pinterest.

YouTube [Click here to watch](https://youtu.be/cqWXGVoDX8Q) Audio [Click here to listen](https://anchor.fm/andreaskayy/episodes/069-Data-Engineering-At-Pinterest-Case-Study-e45ioh)

Table 31.8: Podcast: 069 Engineering Culture At Pinterest

#### Slides:

h[ttps://www.slideshare.net/ConfluentInc/pinterests-](http://www.slideshare.net/ConfluentInc/pinterests-story-of-streaming-hundreds-of-terabytes-of-pins-)story-[of-streaming-](http://www.slideshare.net/ConfluentInc/pinterests-story-of-streaming-hundreds-of-terabytes-of-pins-)h[undreds-of-terabytes-of-pins-](http://www.slideshare.net/ConfluentInc/pinterests-story-of-streaming-hundreds-of-terabytes-of-pins-) ref=h[ttps://www](http://www.confluent.io/kafka-summit-sf18/pinterests-story-of-streaming-hundreds-of-terabytes).confl[uent.io/kafka-summit-sf18/pinterests-story-of-streaming-](http://www.confluent.io/kafka-summit-sf18/pinterests-story-of-streaming-hundreds-of-terabytes)h[undreds-of-terabytes](http://www.confluent.io/kafka-summit-sf18/pinterests-story-of-streaming-hundreds-of-terabytes)

h[ttps://www.slideshare.net/ConfluentInc/building-](http://www.slideshare.net/ConfluentInc/building-pinterest-realtime-ads-platform-using-kafka-stream)pin[terest-realtime-](http://www.slideshare.net/ConfluentInc/building-pinterest-realtime-ads-platform-using-kafka-stream)ads-[platform-using-kafka-stream](http://www.slideshare.net/ConfluentInc/building-pinterest-realtime-ads-platform-using-kafka-stream) ref=h[ttps://www](http://www.confluent.io/kafka-summit-sf18/building-pinterest-real-time-ads-platform-using-kafk).confl[uent.io/kafka-summit-sf18/building-pinterest-real-time-ads-platform-using-](http://www.confluent.io/kafka-summit-sf18/building-pinterest-real-time-ads-platform-using-kafk)k[afk](http://www.confluent.io/kafka-summit-sf18/building-pinterest-real-time-ads-platform-using-kafk)

[https://medium.com/@Pinterest Engineering/building-a-real-time-user-action-counting-system-for-ad](https://medium.com/%40Pinterest_Engineering/building-a-real-time-user-action-counting-system-for-ads-88a60d9c9a)

[https://medium.com/pinterest-engineering/goku-building-a-scalable-and-high-performant-time-series-](https://medium.com/pinterest-engineering/goku-building-a-scalable-and-high-performant-time-series-database-system-a8ff5758a181)

<https://medium.com/pinterest-engineering/building-a-dynamic-and-responsive-pinterest-7d410e99f0a9>

[https://medium.com/@Pinterest Engineering/building-pin-stats-25ec8460e924](https://medium.com/%40Pinterest_Engineering/building-pin-stats-25ec8460e924)

[https://medium.com/@Pinterest Engineering/improving-hbase-backup-efficiency-at-pinterest-86159da](https://medium.com/%40Pinterest_Engineering/improving-hbase-backup-efficiency-at-pinterest-86159da4b954)

[https://medium.com/@Pinterest Engineering/pinterest-joins-the-cloud-native-computing-foundation-](https://medium.com/%40Pinterest_Engineering/pinterest-joins-the-cloud-native-computing-foundation-e3b3e66cb4f)

[https://medium.com/@Pinterest Engineering/using-kafka-streams-api-for-predictive-budgeting-9f58d2](https://medium.com/%40Pinterest_Engineering/using-kafka-streams-api-for-predictive-budgeting-9f58d206c996)

[https://medium.com/@Pinterest Engineering/auto-scaling-pinterest-df1d2beb4d64](https://medium.com/%40Pinterest_Engineering/auto-scaling-pinterest-df1d2beb4d64)

## Data Science @Salesforce

[https://engineering.salesforce.com/building-a-scalable-event-pipeline-with-heroku-and-salesforce-2549](https://engineering.salesforce.com/building-a-scalable-event-pipeline-with-heroku-and-salesforce-2549cb20ce06)

## Data Science @Siemens Mindsphere

**Podcast Episode:** #059 What Is The Siemens Mindsphere IoT Platform?

The Internet of things is a huge deal. There are many platforms available. But, which one is actually good? Join me on a 50 minute dive into the Siemens Mind- sphere online documentation. I have to say I was super unimpressed by what I found. Many limitations, unclear architecture and no pricing available? Not good!

YouTube [Click here to watch](https://youtu.be/HEd5Tsuy5HE) Audio [Click here to listen](https://anchor.fm/andreaskayy/episodes/059-A-Look-Into-The-Siemens-Mindsphere-IoT-Platform---059-e45ihn)

Table 31.9: Podcast: 059 What Is The Siemens Mindsphere IoT Platform?

## Data Science @Slack

<https://speakerdeck.com/vananth22/streaming-data-pipelines-at-slack>

## Data Science @Spotify

**Podcast Episode:** #071 Data Engineering At Spotify Case Study

In this episode we are looking at the data engineering at Spotify, my favorite music streaming service. How do they process all that data?

YouTube [Click here to watch](https://youtu.be/0WJZ5wtQRWI) Audio [Click here to listen](https://anchor.fm/andreaskayy/episodes/071-Data-Engineering-At-Spotify-Case-Study-e45iq1)

Table 31.10: Podcast: 071 Data Engineering At Spotify Case Study

#### Slides:

<https://labs.spotify.com/2016/02/25/spotifys-event-delivery-the-road-to-the-cloud-part-i/> <https://labs.spotify.com/2016/03/03/spotifys-event-delivery-the-road-to-the-cloud-part-ii/> <https://labs.spotify.com/2016/03/10/spotifys-event-delivery-the-road-to-the-cloud-part-iii/> <https://www.slideshare.net/InfoQ/scaling-the-data-infrastructure-spotify> <https://www.datanami.com/2018/05/16/big-data-file-formats-demystified/> <https://labs.spotify.com/2017/04/26/reliable-export-of-cloud-pubsub-streams-to-cloud-storage/> <https://labs.spotify.com/2017/11/20/autoscaling-pub-sub-consumers/>

## Data Science @Symantec

<https://www.slideshare.net/planetcassandra/symantec-cassandra-data-modelling-techniques-in-action>

## Data Science @Tinder

[https://www.slideshare.net/databricks/scalable-monitoring-using-apache-spark-and-friends-with-utka](https://www.slideshare.net/databricks/scalable-monitoring-using-apache-spark-and-friends-with-utkarsh-bhatnagar)

## Data Science @Twitter

#### Slides:

<https://www.slideshare.net/sawjd/real-time-processing-using-twitter-heron-by-karthik-ramasamy> [https://www.slideshare.net/sawjd/big-data-day-la-2016-big-data-track-twitter-heron-scale-karthik-ra](https://www.slideshare.net/sawjd/big-data-day-la-2016-big-data-track-twitter-heron-scale-karthik-ramasamy-engineering-manager-twitter)

**Podcast Episode:** #072 Data Engineering At Twitter Case Study

How is Twitter doing Data Engineering? Oh man, they have a lot of cool things to share these tweets.

YouTube [Click here to watch](https://youtu.be/UkqSR3IeLZ8) Audio [Click here to listen](https://anchor.fm/andreaskayy/episodes/072-Data-Engineering-At-Twitter-Case-Study-e45iqq)

Table 31.11: Podcast: 072 Data Engineering At Twitter Case Study

<https://techjury.net/stats-about/twitter/> <https://developer.twitter.com/en/docs/tweets/post-and-engage/overview> <https://www.slideshare.net/prasadwagle/extracting-insights-from-data-at-twitter>

[https://blog.twitter.com/engineering/en us/topics/insights/2018/twitters-kafka-adop](https://blog.twitter.com/engineering/en_us/topics/insights/2018/twitters-kafka-adoption-story.html)tion-story. [html](https://blog.twitter.com/engineering/en_us/topics/insights/2018/twitters-kafka-adoption-story.html)

[https://blog.twitter.com/engineering/en us/topics/infrastructure/2017/the-infrastructure-](https://blog.twitter.com/engineering/en_us/topics/infrastructure/2017/the-infrastructure-behind-twitter-scale.html)behind-twit [html](https://blog.twitter.com/engineering/en_us/topics/infrastructure/2017/the-infrastructure-behind-twitter-scale.html)

[https://blog.twitter.com/engineering/en us/topics/infrastructure/2019/the-start-of-a-](https://blog.twitter.com/engineering/en_us/topics/infrastructure/2019/the-start-of-a-journey-into-the-cloud.html)journey-into-the [html](https://blog.twitter.com/engineering/en_us/topics/infrastructure/2019/the-start-of-a-journey-into-the-cloud.html)

<https://www.slideshare.net/billonahill/twitter-heron-in-practice> <https://streaml.io/blog/intro-to-heron> <https://www.youtube.com/watch?v=3QHGhnHx5HQ> [https://hbase.apache.org](https://hbase.apache.org/)

[https://db-engines.com/en/system/Amazon+DynamoDB%3BCassandra%3BGo](https://db-engines.com/en/system/Amazon%2BDynamoDB%3BCassandra%3BGoogle%2BCloud%2BBigtable%3BHBase)ogle+Cloud+ [Bigtable%3BHBase](https://db-engines.com/en/system/Amazon%2BDynamoDB%3BCassandra%3BGoogle%2BCloud%2BBigtable%3BHBase)

## Data Science @Uber

<https://eng.uber.com/uber-big-data-platform/> <https://eng.uber.com/aresdb/> <https://www.uber.com/us/en/uberai/>

## Data Science @Upwork

[https://www.slideshare.net/databricks/how-to-rebuild-an-endtoend-ml-pipeline-with-databricks-and-](https://www.slideshare.net/databricks/how-to-rebuild-an-endtoend-ml-pipeline-with-databricks-and-upwork-with-thanh-tran)

## Data Science @Woot

[https://aws.amazon.com/de/blogs/big-data/our-data-lake-story-how-woot-com-built-a-serverless-data](https://aws.amazon.com/de/blogs/big-data/our-data-lake-story-how-woot-com-built-a-serverless-data-lake-on-aws/)

## Data Science @Zalando

**Podcast Episode:** #087 Data Engineering At Zalando Case Study Talk

I had a great conversation about data engineering for online retailing with Michal Gancarski and Max Schultze. They showed Zalando’s data platform and how they build data pipelines. Super interesting especially for AWS users.

YouTube [Click here to watch](https://youtu.be/IXOLsNA6Hm0)

Table 31.12: Podcast: 087 Data Engineering At Zalando Case Study

Do me a favor and give these guys a follow on LinkedIn: LinkedIn of Michal: [https:](https://www.linkedin.com/in/michalgancarski/)

[//www.linkedin.com/in/michalgancarski/](https://www.linkedin.com/in/michalgancarski/)

LinkedIn of Max: <https://www.linkedin.com/in/max-schultze-b11996110/>

Zalando has a tech blog with more infos and there is also a meetup in Berlin: Zalando Blog: <https://jobs.zalando.com/tech/blog/>

Next Zalando Data Engineering Meetup: [https://www.meetup.com/Zalando-Tech-Even](https://www.meetup.com/Zalando-Tech-Events-Berlin/events/262032282/)ts-Berlin/ [events/262032282/](https://www.meetup.com/Zalando-Tech-Events-Berlin/events/262032282/)

Interesting tools: AWS CDK: [https://docs.aws.amazon.com/cdk/latest/guide/what-is.](https://docs.aws.amazon.com/cdk/latest/guide/what-is.html) [html](https://docs.aws.amazon.com/cdk/latest/guide/what-is.html)

Delta Lake: <https://delta.io/>

AWS Step Functions: https://aws.amazon.com/step-functions/AWSStateLanguage:https:

//states-language.net/spec.html

Youtube channel of the meetup: [https://www.youtube.com/channel/UCxwul7aBm2LybbpK](https://www.youtube.com/channel/UCxwul7aBm2LybbpKGbCOYNA/playlists%20talk%20at%20Spark%2BAI)GbCOYNA [playliststalkatSpark+AI](https://www.youtube.com/channel/UCxwul7aBm2LybbpKGbCOYNA/playlists%20talk%20at%20Spark%2BAI)

Summit about Zalando’s Processing Platform: [https://databricks.com/session/continuous-applications-](https://databricks.com/session/continuous-applications-at-scale-of-100-teams-with-databricks-delta-and-structured-streaming)

Talk at Strata London slides: [https://databricks.com/session/continuous-applications-at-scale-of-100-t](https://databricks.com/session/continuous-applications-at-scale-of-100-teams-with-databricks-delta-and-structured-streaming)

[https://jobs.zalando.com/tech/blog/what-is-hardcore-data-science--in-practice/?gh src=](https://jobs.zalando.com/tech/blog/what-is-hardcore-data-science--in-practice/?gh_src=4n3gxh1) [4n3gxh1](https://jobs.zalando.com/tech/blog/what-is-hardcore-data-science--in-practice/?gh_src=4n3gxh1)

[https://jobs.zalando.com/tech/blog/complex-event-generation-for-business-process-monitoring-using-](https://jobs.zalando.com/tech/blog/complex-event-generation-for-business-process-monitoring-using-apache-flink/)

# Part V

**1001 Data Engineering Interview Questions**

Looking for a job or just want to know what people find important? In this chapter you can find a lot of interview questions we collect on the stream.

Ultimately this should reach at least one thousand and one questions.

**But Andreas, where are the answers??** Answers are for losers. I have been thinking a lot about this and the best way for you to prepare and learn is to look into these questions yourself.

This cookbook or Google will help you a long way. Some Questions we discuss directly on the live stream.

# Live Streams

First live stream where we started to collect these questions.

**Podcast Episode:** #096 1001 Data Engineering Interview Questions

First live stream where we collect and try to answer as many interview questions as possible. If this helps people and is fun we do this regularly until we reached 1000 and one.

YouTube [Click here to watch](https://youtu.be/WbqRH2r3N40)

Table 32.1: Podcast: 096 1001 Data Engineering Interview Questions

# All Interview Questions

The interview questions are roughly structured like the sections in the ”Basic data Engi- neering Skills” part. This makes it easier to navigate this document. I still need to sort them accordingly.

## SQL DBs

* What are windowing functions?
* What is a stored procedure
* Why would you use them?
* What are atomic attributes
* Explain ACID props of a database
* How to optimize queries
* What are the different types of JOIN (CROSS, INNER, OUTER)
* What is the difference between Clustered Index and Non-Clustered Index - with examples?

## The Cloud

* What is serverless
* What’s the difference between IaaS, PaaS and SaaS
* How do you move from the ingest layer to the Cosumption layer? (In Serverless)
* Whats the difference between cloud and edge and on-premises
* What is edge computing

## Linux

* What is crontab

## Big Data

* What are the 4 V’s
* Which one is most important?

## Kafka

* What is a topic
* How to ensure FIFO
* How do you know if all messages in a topic have been fully consumed
* What are brokers
* What are consumergroups
* What is a producer

## Coding

* What’s the difference between an object and a class
* Explain immutability
* What are AWS Lambda functions and why would you use them
* Difference between library, framework and package
* How to reverse a linked list
* difference between args and kwargs
* Difference between oop and functional programming

## NoSQL DBs

* What’s a key/value (rowstore) store
* What’s a columnstore
* Diff between Row an col.store
* What’s a document store
* Difference between Redshift and Snowflake

## Hadoop

* What File Formats can you use in Hadoop
* Whats the difference between a name and a datanode
* What is HDFS
* What is the purpous of YARN

## Lambda Architecture

* what is streaming and batching
* what is the upside of streamtin vs batching
* What’s the difference between lambda and kappa architecture
* Can you sync the batch and streaming layer and if yes how

## Python

* Difference between list tuples and dictionary

## Data Warehouse & Data Lake

* What is a data lake?
* What is a data warehouse
* Are there data lake warehouses?
* Two Datalakes within single warehouse?
* What is a data maart?
* what is a slow changing dimension (types)
* What is a surrogate key and why use them?

## APIs (REST)

* What does REST mean?
* What is idempotency
* What are common REST API frameworks (Jersey and Spring)

## Apache Spark

* What’s an RDD
* What is a dataframe
* What is a dataset
* How is a dataset typesafe
* What is Parquet
* What’s Avro
* Difference between Parquet and Avro
* Tumbling Windows Vs. Sliding Windows
* Difference between batch ans stream processing
* What are microbatches

## MapReduce

* What’s a use case of mapreduce
* Write a pseudo code for Wordcount
* What is a combiner

## Docker & Kubernetes

* What is a container
* Difference between Docker Container and a Virtual PC
* What s the easiest way to learn kubernetes fast

## Data Pipelines

* What is an example of a serverless pipeline
* What’s difference between at most once vs at least once vs exactly once
* What systems provide transactions
* What is a ETL pipeline

## Airflow

* What is a DAG (in context of airflow/luigi)
* What are Hooks/ is a hook
* What are Operators
* How to branch?

## DataViszualization

* What’s a BI tool

## Security/Privacy

* What is Kerberos
* What is a firewall
* Whats GDPR?
* What’s anonymization

## Distrubuted Systems

* how clusters reach consensus (the answer was using consensus protocols like Paxos or Raft). Good I didnt have to explain paxos
* What is the cap theorem / explain it (What factors should be considered when choosing a DB?)
* How to choose right storage for different data consumers? It’s always a tricky question

## Apache Flink

* what is Flink used for
* Flink vs Spark?

## GitHub

* What are branches
* What are commits
* What’s a pull request

## Dev/Ops

* What is continuous integration
* What is continuous deployment
* Difference CI/CD

## Development / Agile

* What is Scrum
* What is OKR
* What is Jira and what is it used for

# Bibliography

[1] J. Ely and I. Stavrov1, *Analyzing chalk dust and writing speeds: computational and geometric approaches*, BoDine Journal of Mathematics **3** (2001), 14-159.

# List of Figures

2.1 The Machine Learning Pipeline [13](#_bookmark5)

* 1. Common SQL Platform Architecture [35](#_bookmark62)
  2. Scaling up a SQL Database [36](#_bookmark64)
  3. Scaling out a SQL Database [37](#_bookmark66)

13.1 Platfrom Blueprint [39](#_bookmark69)

* 1. Batch Processing Pipeline [43](#_bookmark77)
  2. Stream Processing Pipeline [44](#_bookmark79)
  3. Hadoop Ecosystem Components [48](#_bookmark92)
  4. Connections between tools [49](#_bookmark94)
  5. Flume Integration [50](#_bookmark95)
  6. HDFS Master and Data Nodes [60](#_bookmark129)
  7. Distribution of Blocks for a 512MB File [60](#_bookmark130)
  8. Mapping of input files and reducing of mapped records [68](#_bookmark152)
  9. MapReduce Example of Time Series Data [70](#_bookmark155)
  10. The Map Reduce Process [71](#_bookmark157)
  11. Hadoop vs Spark capabilities [72](#_bookmark161)
  12. Spark Using Hadoop Data Locality [75](#_bookmark170)
  13. Spark Resource Management With YARN [77](#_bookmark181)
  14. Old Netflix Batch Processing Pipeline [102](#_bookmark269)
  15. Netflix Trending Now Feature [103](#_bookmark270)
  16. Netflix Streaming Pipeline [104](#_bookmark271)

**List of Tables**

|  |  |  |  |
| --- | --- | --- | --- |
| 2.1 | Podcast: | 050 Data Engineer Scientist or Analyst Which One Is For You? | [12](#_bookmark3) |
| 2.2 | Podcast: | 048 From Wannabe Data Scientist To Engineer My Journey . . | [14](#_bookmark7) |
| 5.1 | Podcast: | 070 Engineering Culture At Spotify . . . . . . . . . . . . . . . . | [22](#_bookmark23) |
| 10.1 | Podcast: | 082 Reading Tweets With Apache Nifi & IaaS vs PaaS vs SaaS | [30](#_bookmark46) |
| 10.2 | Podcast: | 076 Cloud vs On-Premise . . . . . . . . . . . . . . . . . . . . . | [30](#_bookmark49) |
| 14.1 | Podcast: | 077 Lambda Architecture and Kappa Architecture . . . . . . . | [43](#_bookmark75) |
| 14.2 | Podcast: | 066 How To Do Data Science From A Data Engineers.. . . . . . | [45](#_bookmark85) |
| 15.1 | Podcast: | 055 Data Warehouse vs Data Lake . . . . . . . . . . . . . . . . | [46](#_bookmark87) |
| 16.1 | Podcast: | 060 What Is Hadoop And Is Hadoop Still Relevant In 2019? . . | [47](#_bookmark90) |
| 18.1 | Podcast: | 033 How APIs Rule The World . . . . . . . . . . . . . . . . . . | [56](#_bookmark113) |
| 18.2 | Podcast: | 081 Twitter API Research . . . . . . . . . . . . . . . . . . . . . | [56](#_bookmark115) |
| 19.1 | Podcast: | 056 NoSQL Key Value Stores Explained with HBase . . . . . . | [59](#_bookmark128) |
| 19.2 | Podcast: | 093 What is MongoDB . . . . . . . . . . . . . . . . . . . . . . . | [61](#_bookmark132) |
| 19.3 | Podcast: | What is Elasticsearch & Why is It So Popular? . . . . . . . . . | [62](#_bookmark134) |
| 19.4 | Podcast: | Druid NoSQL DB And Analytics DB Introduction . . . . . . . | [63](#_bookmark139) |
| 20.1 | Podcast: | 039 Is ETL Dead for Data Science and Big Data? . . . . . . . . | [65](#_bookmark144) |
| 22.1 | Podcast: | Machine Learning In Production . . . . . . . . . . . . . . . . . | [80](#_bookmark190) |
| 25.1 | Podcast: | 068 How to Build a Budget Data Science PC . . . . . . . . . . | [89](#_bookmark222) |
| 26.1 | Podcast: | 081 Twitter API Research . . . . . . . . . . . . . . . . . . . . . | [90](#_bookmark224) |
| 27.1 | Podcast: | 082 Reading Tweets With Apache Nifi . . . . . . . . . . . . . . | [91](#_bookmark226) |
| 27.2 | Podcast: | 085 Trying to read Tweets with Nifi Part 2 . . . . . . . . . . . | [91](#_bookmark227) |
| 28.1 | Podcast: | 086 How to Write from Nifi to Kafka Part 1 . . . . . . . . . . . | [92](#_bookmark229) |
| 28.2 | Podcast: | 088 How to Write from Nifi to Kafka Part 2 . . . . . . . . . . . | [92](#_bookmark230) |

|  |  |  |  |
| --- | --- | --- | --- |
| 31.1 | Podcast: | 063 Data Engineering At Airbnb Case Study . . . . . . . . . . | [96](#_bookmark243) |
| 31.2 | Podcast: | 064 Data Engineering At Booking.com Case Study . . . . . . . | [97](#_bookmark249) |
| 31.3 | Podcast: | 065 Data Engineering At CERN Case Study . . . . . . . . . . . | [98](#_bookmark251) |
| 31.4 | Podcast: | 073 Data Engineering At LinkedIn Case Study . . . . . . . . . | [100](#_bookmark263) |
| 31.5 | Podcast: | 067 Data Engineering At NASA Case Study . . . . . . . . . . . | [101](#_bookmark266) |
| 31.6 | Podcast: | 062 Data Engineering At Netflix Case Study . . . . . . . . . . . | [101](#_bookmark268) |
| 31.7 | Podcast: | 083 Data Engineering at OLX Case Study . . . . . . . . . . . . | [105](#_bookmark273) |
| 31.8 | Podcast: | 069 Engineering Culture At Pinterest . . . . . . . . . . . . . . . | [105](#_bookmark277) |
| 31.9 | Podcast: | 059 What Is The Siemens Mindsphere IoT Platform? . . . . . . | [106](#_bookmark280) |

* 1. Podcast: 071 Data Engineering At Spotify Case Study [107](#_bookmark283)
  2. Podcast: 072 Data Engineering At Twitter Case Study [108](#_bookmark287)
  3. Podcast: 087 Data Engineering At Zalando Case Study [109](#_bookmark292)

32.1 Podcast: 096 1001 Data Engineering Interview Questions [113](#_bookmark295)