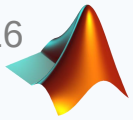
Analysis of Operational Flight Data in Hadoop using MapReduce and the MATLAB Distributed Computing Server (MDCS)

MATLAB EXPO 2016

Lukas Höhndorf
Javensius Sembiring, Robin Karpstein, and Florian Holzapfel



Outline

- Mission of the Flight Safety Group
- Big Data Concepts Available in MATLAB
- MATLAB Distributed Computing Server (MDCS)
- Application
- Summary

Background

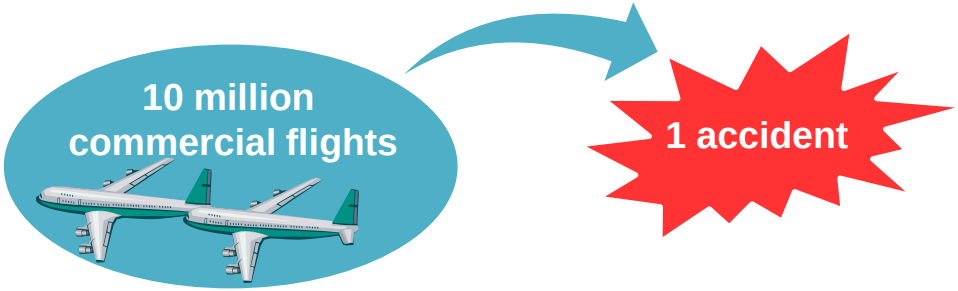
1

ICAO DOC 9859

- Airlines are required to implement a Safety Management System (SMS)
- SMS requires operators also to define their own **Acceptable Level of Safety (ALoS).**

Definition of ALoS within ICAO DOC 9859:
*"The **minimum level of safety** performance […] of a service provider, **as defined in its safety management** […] ."*

- **Europe** aims at less than one accident per ten million flights (i.e. **accident probability of 10^{-7} per flight**).



2

Flightpath 2050

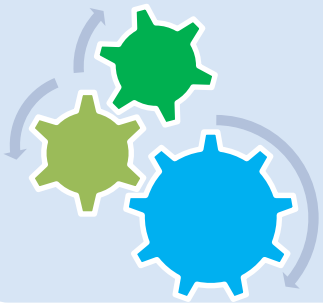
General Concept

$$P_{Accident} = \frac{Number\ of\ accidents}{Number\ of\ flights}$$

Classical statistical approach

$$P_{Accident} = \frac{0}{400\,000} = 0$$

Runway overrun example



Predictive Analysis:
Making **quantitative statements** about the future state based on **previous experience and knowledge**.

General Concept

Contributing Factors

- Weight
- Wind
- Speed
- Flaps
- Start of Braking
- …
- …
- …

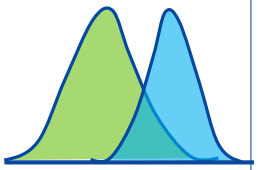
Model and Method

Incident Model

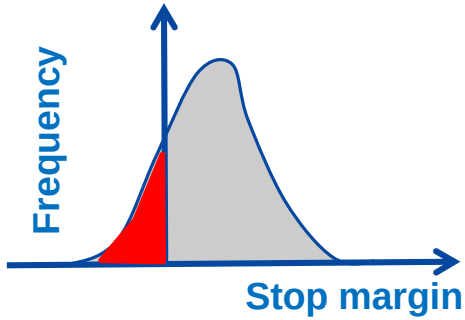
e.g. Overrun

Image: IATA

Advanced Statistical Methods



Goal



Incident Probability
i.e. Overrun Probability

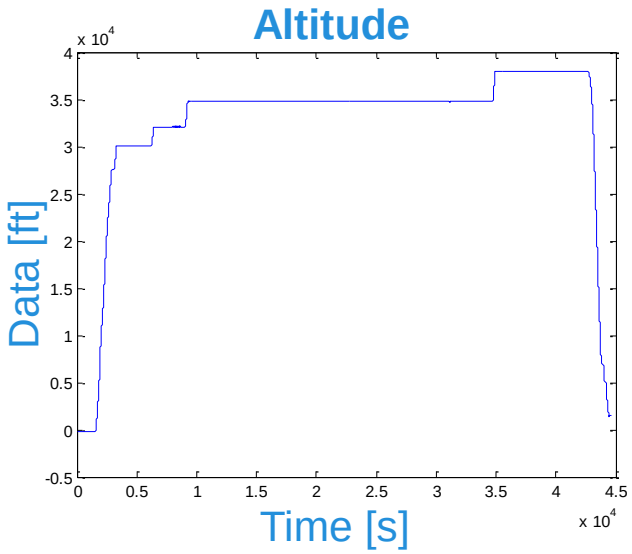
General Concept

General Concept

To quantify accident probabilities in aviation, there are a lot of computations and interactions with a huge volume and different type of data necessary.

Flight Data

- Up to 2500 variables (depending on the aircraft and the airline) are recorded.
- Frequency usually between ¼ Hertz and 8 Hertz, depending on the variable.
- Number of the recorded variables increased significantly in the last years.



Source: http://www.sagem.com/aerospace/commercial-aircraft/information-system/aircraft-condition-monitoring-system-acms

Big Data

- Nowadays, many people talk about Big Data.

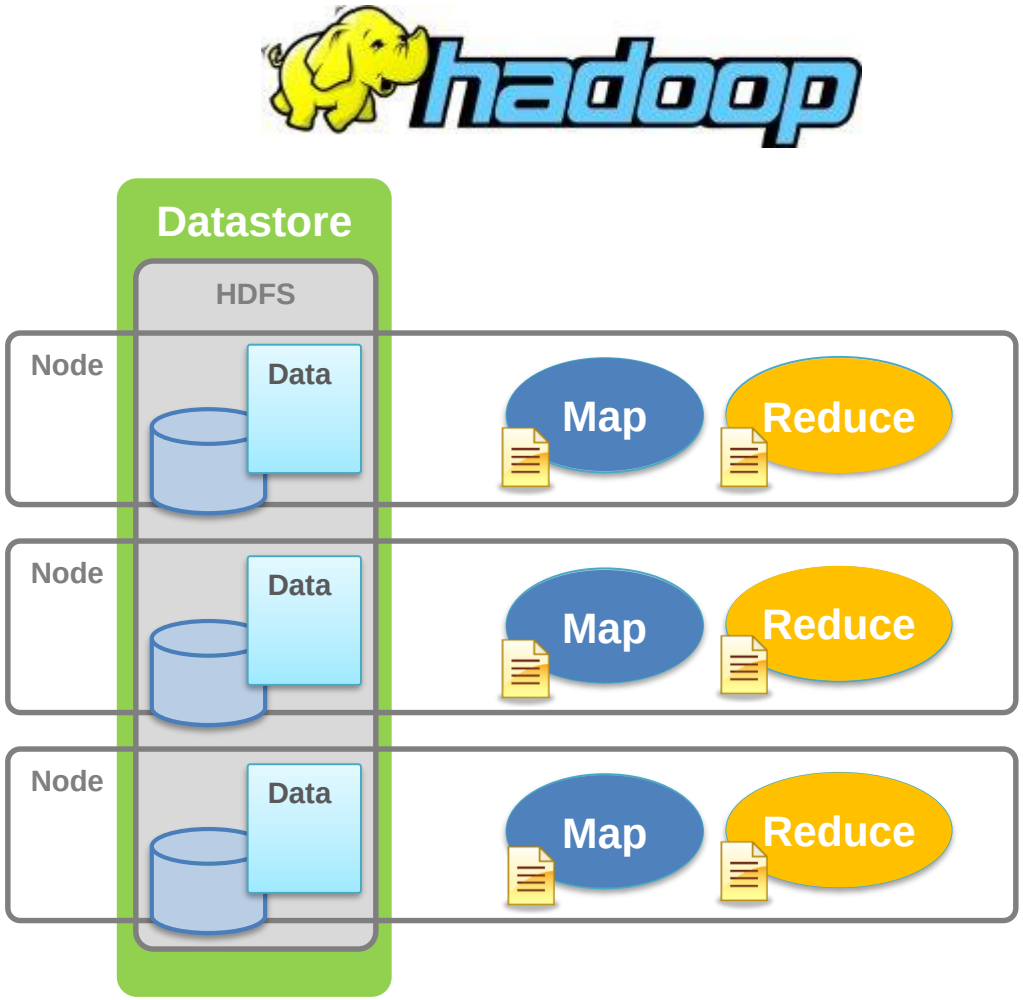
  To describe Big Data, the 4 V's are common:

  - Volume
  - Velocity
  - Variety
  - Veracity

Source: https://www.ucl.ac.uk/big-data/bdi/images/data-head.jpg

- Goal of the project:
  - Learn about existing Big Data concepts available in MATLAB
  - Apply these concepts to flight data analysis

Hadoop



Source: MATHWORKS

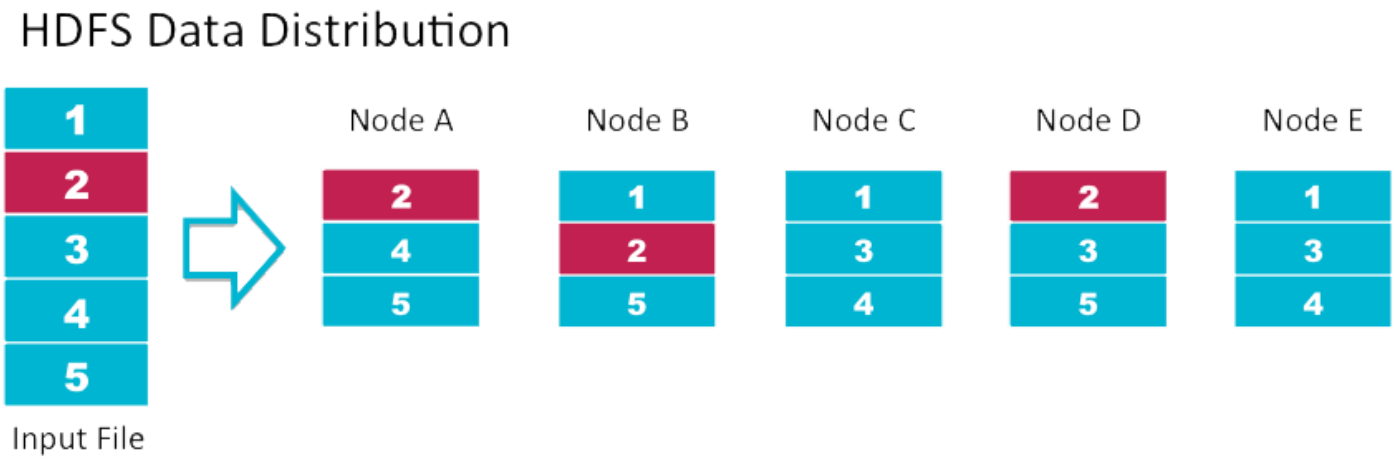
Hadoop - Ecosystem

Additional software packages for Hadoop:

- Flume - Efficiently collecting, aggregating, and moving large amounts of log data
- Sqoop - Transfering data between relational databases and Hadoop
- Hbase - Open source, non-relational, distributed database
- Hive - Data Warehouse for providing data summarization, query, and analysis
- Oozie - Server-based workflow scheduling system to manage Hadoop jobs
- Pig - High-level platform for creating MapReduce programs used with Hadoop
- Spark - Open source cluster computing framework
- …

HDFS

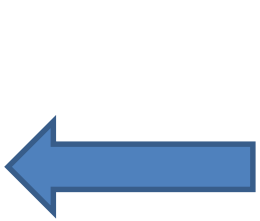
- A distributed file system designed to run on commodity hardware.
- HDFS provides high throughput access to application data and is suitable for applications that have large data sets.
- Files are cut apart in chunks and distributed among various data nodes.

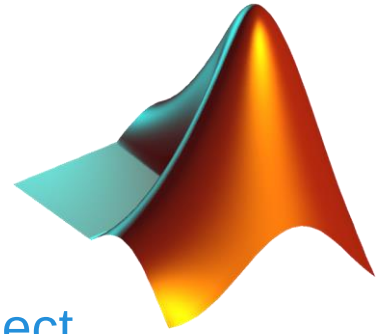


Source: http://www.cloudera.com/content/dam/cloudera/product-assets/hdfs-data-distribution.png

MATLAB Interface to HDFS

- From MATLAB, direct access to HDFS using datastore objects is very convenient.





MATLAB datastore – Object

Access data with MATLAB read

(observe analogy between HDFS chunk size and MATLAB ReadSize)

Source: http://hadoop.apache.org/

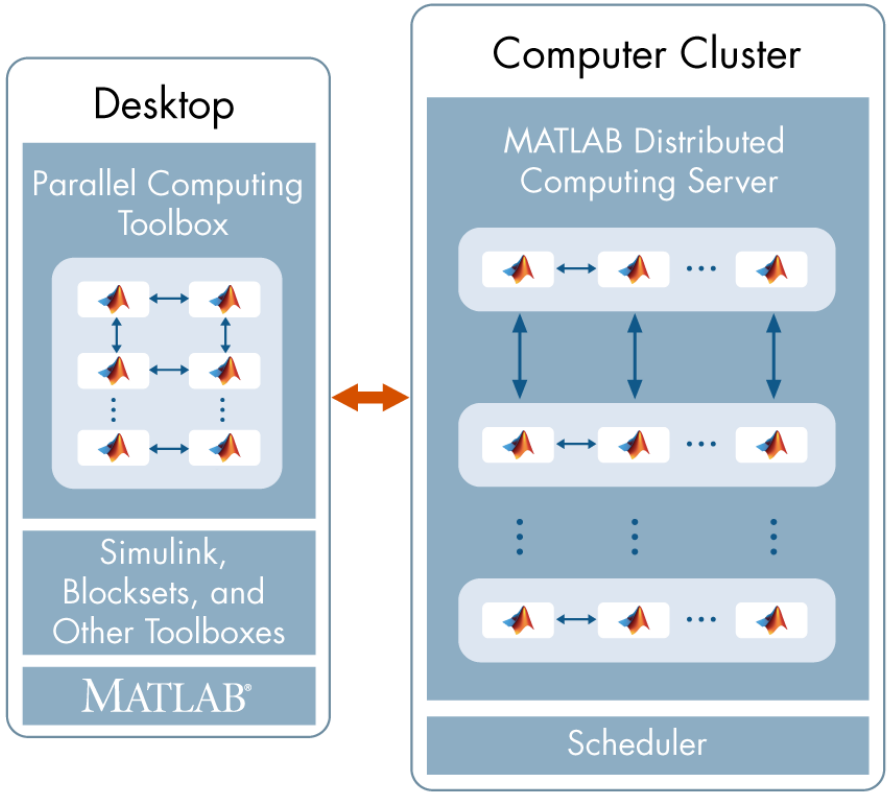
MapReduce

- MapReduce is a programming technique for analyzing data sets that do not fit in memory
- MapReduce consists of two parts: Map and Reduce
- Map: The Map algorithms are applied to individual chunks of the big data file (compare chunk structure of HDFS)
- Reduce: The results off all the applications of Map are combined in an application of a Reduce function.
- Typical example: Finding minimum value of a (very big) array

```
outds = mapreduce(ds,mapfun,reducefun,mr);
```

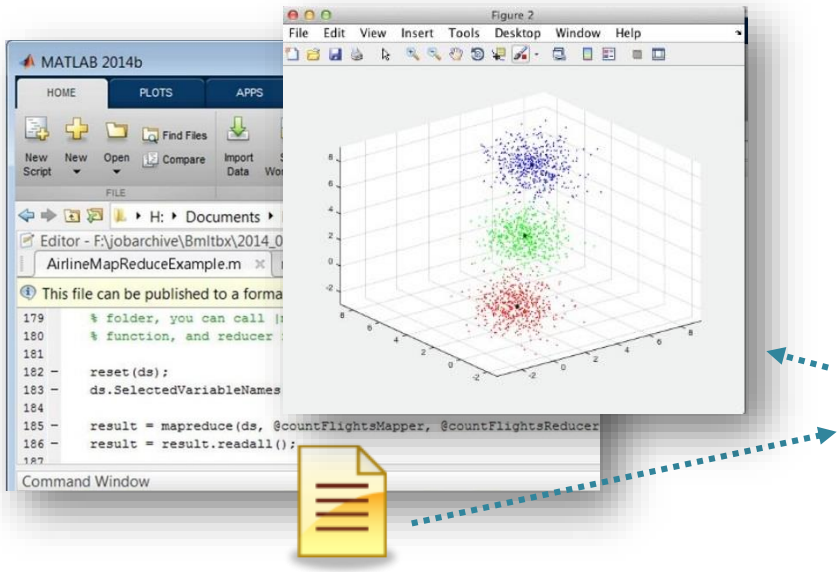
MATLAB Distributed Computing Server (MDCS)

- MDCS lets you run computationally intensive MATLAB programs and Simulink models on computer clusters, clouds, and grids, enabling you to speed up computations and solve large problems.
- Since R2014b, MDCS can be directly connected to Hadoop (using the Hadoop Job Scheduler)
- However, configure the MATLAB Job Scheduler (MJS) proved to be more convenient in our situation.

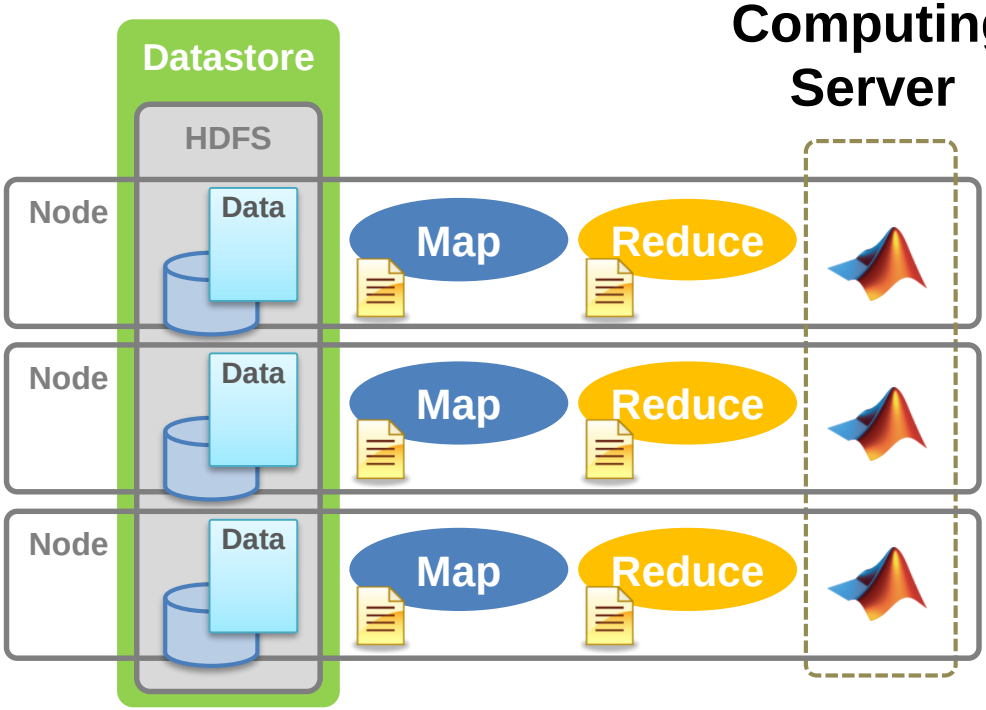


Source: http://www.mathworks.com/help/mdce/index.html, http://de.mathworks.com/cmsimages/62006_wl_mdcs_fig1_wl.png

MATLAB Distributed Computing Server (MDCS)



map.m
reduce.m



Hadoop

Source: MATHWORKS

Application

Given: Flight data as time series and stored in Hadoop (File size depends on many factors; currently up to 5 GB)

Goal: Detect departure and arrival airports and runways using MapReduce and the MDCS

Application

- 1st Step: “Roughly” detect location of lift off and touchdown
  → **Map**
- 2nd Step: Compare coordinates with a navigation database to obtain airports and runways
  → **Reduce**

Application

Air / Ground	Latitude	Longitude
…	…	…
1	48°21'57.62"N	11°48'35.73"E
1	48°21'56.26"N	11°48'4.15"E
1	48°21'53.38"N	11°47'34.21"E
0	48°21'51.12"N	11°47'0.36"E
0	48°21'48.34"N	11°46'27.18"E
0	48°21'46.50"N	11°46'4.12"E
…	…	…

Chunk 1

Chunk 2

Map

Is there a lift off or touchdown in the chunk?

Chunk 1: No

Chunk 2: **Yes**

→

Save Coordinates for Lift Off / Touchdown

Chunk 3 …

Reduce

- Aggregate information from all chunks
- Compare coordinates with navigation database and retrieve airports and runways

departureAirport	departureRunway
EDDM	26R

Results

Following scenarios were tested with MapReduce:

1. Local (Intel Core i5 4x3.1 GHZ, 16 GB DDR RAM, MATLAB R2015a)

2. Hadoop Cluster using Hadoop Job Scheduler (due to complicated configuration of the Hadoop Job Scheduler, MapReduce was conducted on only 1 machine, Intel Core i7 4x3.4 GHZ, 32 GB DDR RAM, MATLAB R2014b)

3. Hadoop Cluster using MATLAB Job Scheduler (whole cluster available, 2 x Intel Core i7 4x3.4 GHZ, 1x Intel Core i7 2x3.4 GHZ, 32/16/4 GB DDR RAM, MATLAB R2014b)

Location	Duration per flight
Local	~ 36 seconds
Hadoop Cluster using Hadoop Job Scheduler	~ 48 seconds
Hadoop Cluster using MATLAB Job Scheduler	~ 6 seconds

Summary

Our Experiences with the Big Data Concepts Available in MATLAB

Big Data Concepts and MATLAB

- Hadoop Storage (HDFS) Beneficial for Storing Huge Data Files
- Hadoop Storage (HDFS) Accessible from MATLAB Using Datastore
- MapReduce Concept Available in MATLAB Allows to Analyze Big Data Files
- Combination of MapReduce and MDCS for Parallelization is Possible

Institute of Flight System Dynamics
Technische Universität München
Boltzmannstraße 15
D-85748 Garching bei München
Deutschland / Germany
Phone: +49 89 289-16080
Fax: +49 89 289-16058

Lukas Höhndorf, lukas.hoehndorf@tum.de
Javensius Sembiring, javensius.sembiring@tum.de
Robin Karpstein, robin.karpstein@tum.de
Florian Holzapfel, florian.holzapfel@tum.de

Ludwig Drees, ludwig.drees@tum.de
Chong Wang, chong.wang@tum.de
Phillip Koppitz, phillip.koppitz@tum.de
Joachim Siegel, joachim.siegel@mytum.de

Thank you for your attention