# Introduction to Big Data

## ECAP456

Dr. Rajni Bhalla

Associate Professor

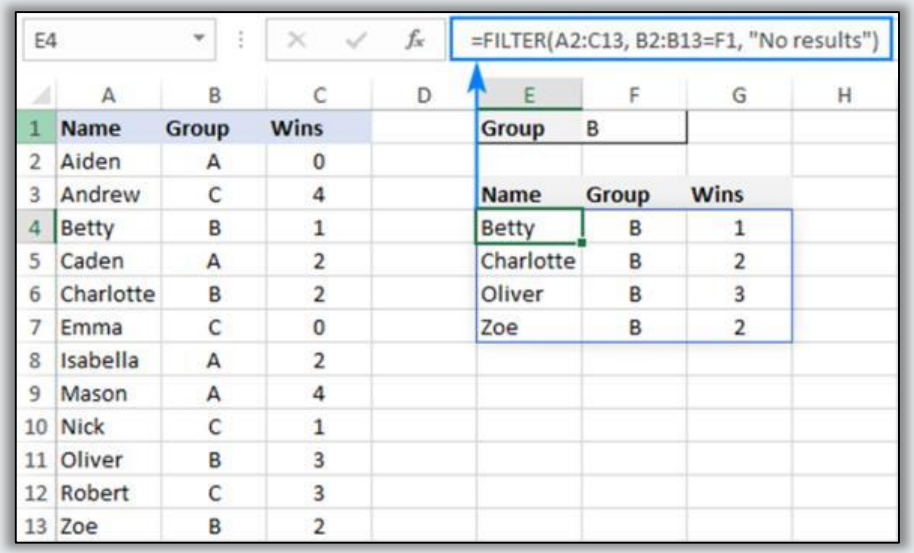
# Learning Outcomes

After this lecture, you will be able to

- learn big data management and processing using datameer

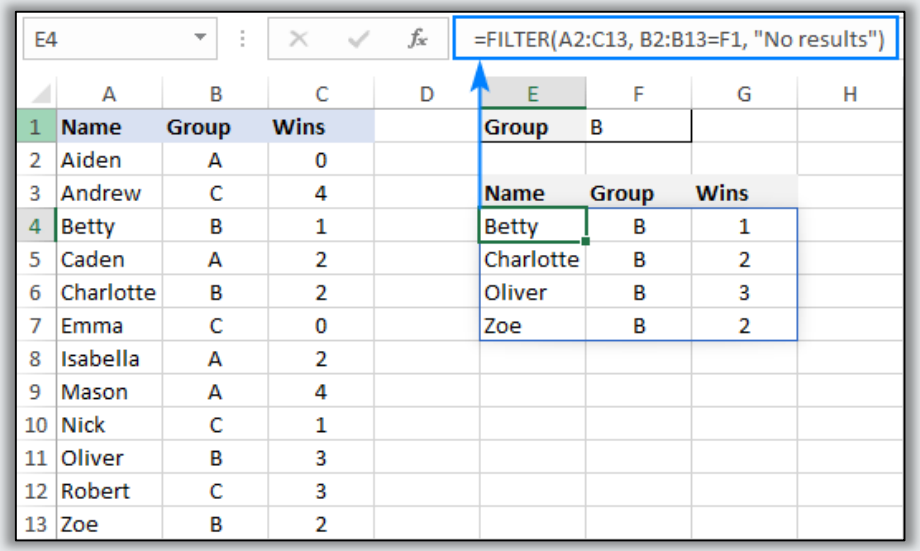
# Introduction

# Introduction



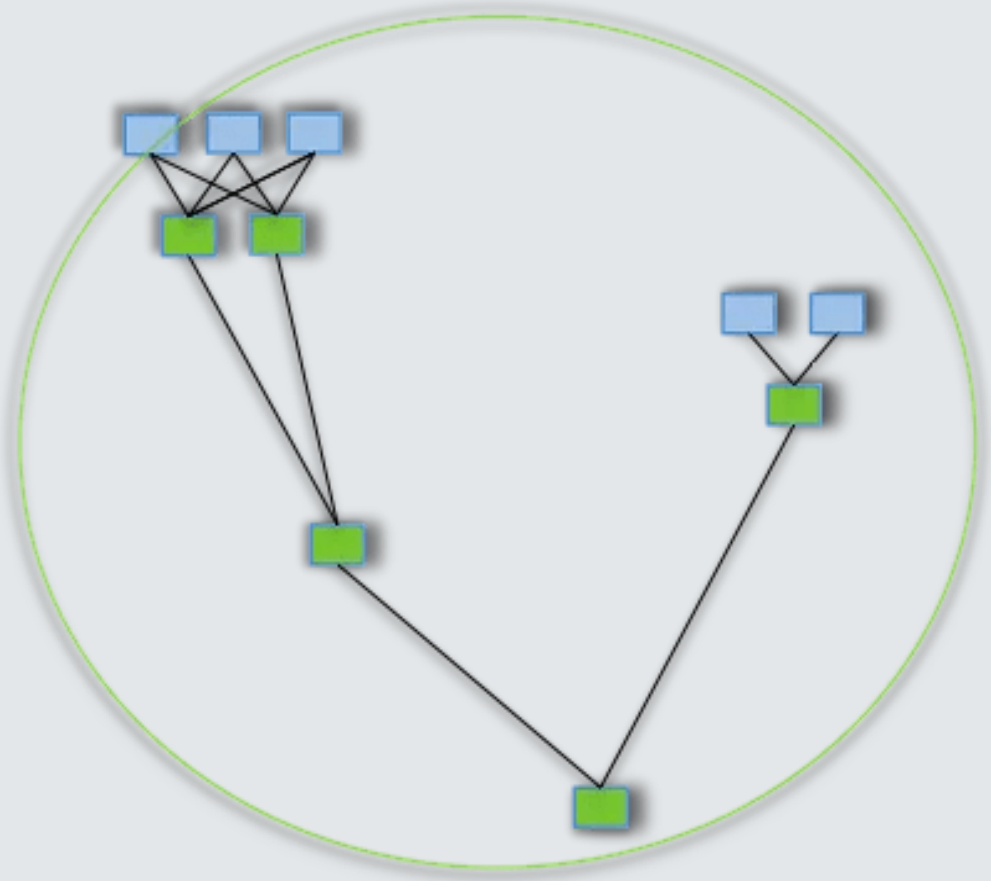

| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | **Name** | **Group** | **Wins** | | **Group** | B | | |
| 2 | Aiden | A | 0 | | | | | |
| 3 | Andrew | C | 4 | | **Name** | **Group** | **Wins** | |
| 4 | Betty | B | 1 | | Betty | B | 1 | |
| 5 | Caden | A | 2 | | Charlotte | B | 2 | |
| 6 | Charlotte | B | 2 | | Oliver | B | 3 | |
| 7 | Emma | C | 0 | | Zoe | B | 2 | |
| 8 | Isabella | A | 2 | | | | | |
| 9 | Mason | A | 4 | | | | | |
| 10 | Nick | C | 1 | | | | | |
| 11 | Oliver | B | 3 | | | | | |
| 12 | Robert | C | 3 | | | | | |
| 13 | Zoe | B | 2 | | | | | |

# Introduction



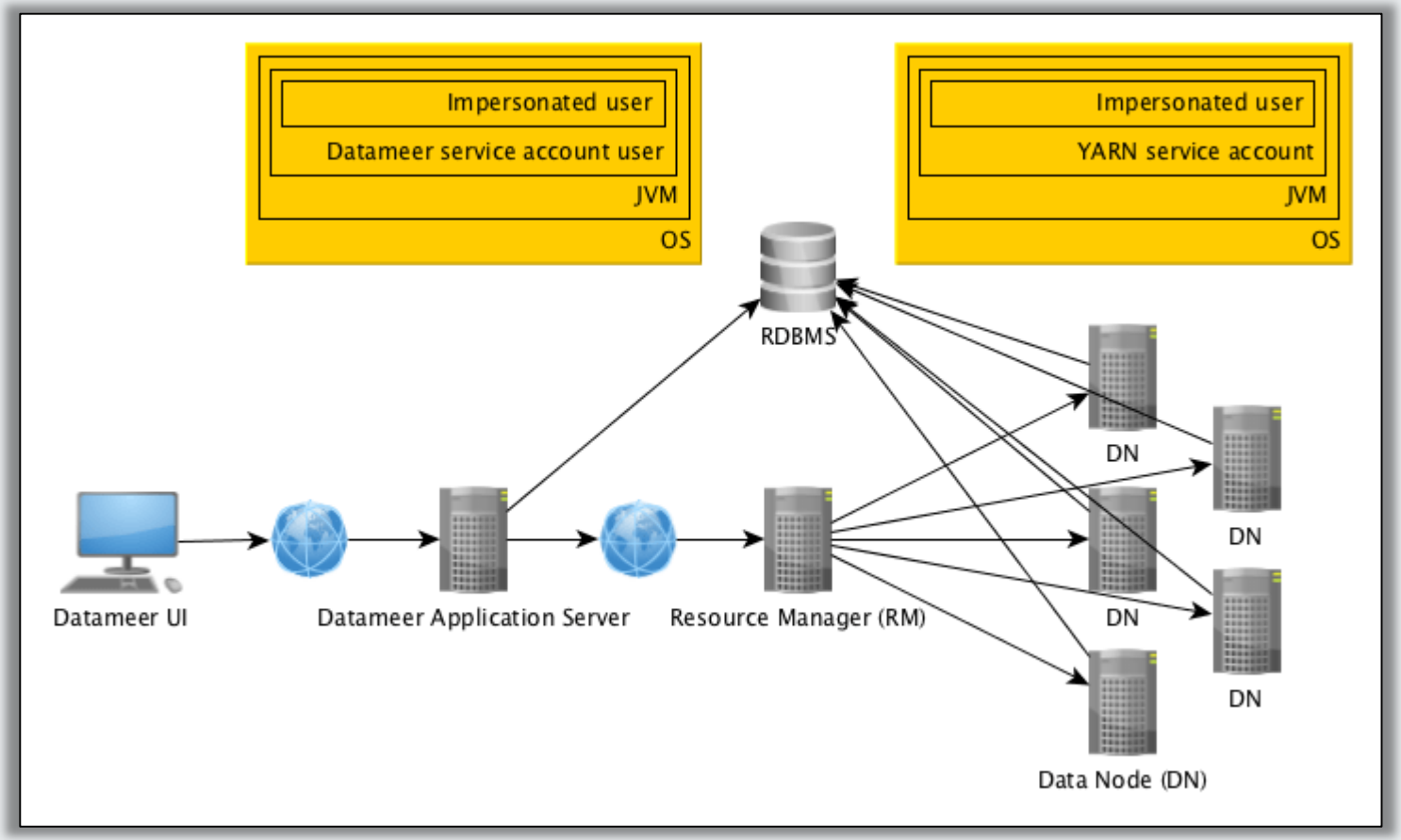

| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | **Name** | **Group** | **Wins** | | **Group** | B | | |
| 2 | Aiden | A | 0 | | | | | |
| 3 | Andrew | C | 4 | | **Name** | **Group** | **Wins** | |
| 4 | Betty | B | 1 | | Betty | B | 1 | |
| 5 | Caden | A | 2 | | Charlotte | B | 2 | |
| 6 | Charlotte | B | 2 | | Oliver | B | 3 | |
| 7 | Emma | C | 0 | | Zoe | B | 2 | |
| 8 | Isabella | A | 2 | | | | | |
| 9 | Mason | A | 4 | | | | | |
| 10 | Nick | C | 1 | | | | | |
| 11 | Oliver | B | 3 | | | | | |
| 12 | Robert | C | 3 | | | | | |
| 13 | Zoe | B | 2 | | | | | |

# Introduction

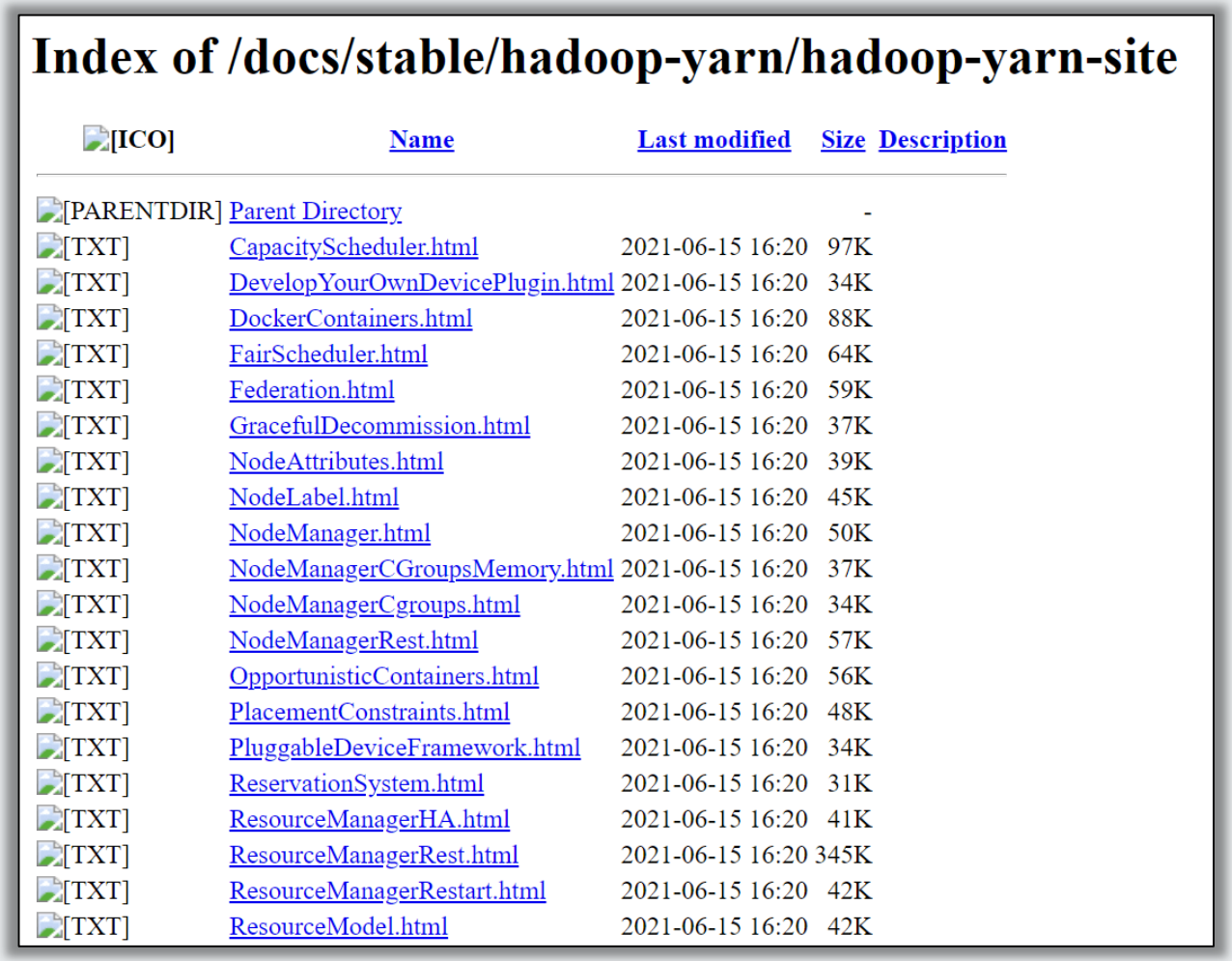



Splitting up workloads into smaller pieces.

# Introduction



# Introduction



## Index of /docs/stable/hadoop-yarn/hadoop-yarn-site

| [ICO] | Name | Last modified | Size | Description |
|---|---|---|---|---|
| [PARENTDIR] | Parent Directory | | - | |
| [TXT] | CapacityScheduler.html | 2021-06-15 16:20 | 97K | |
| [TXT] | DevelopYourOwnDevicePlugin.html | 2021-06-15 16:20 | 34K | |
| [TXT] | DockerContainers.html | 2021-06-15 16:20 | 88K | |
| [TXT] | FairScheduler.html | 2021-06-15 16:20 | 64K | |
| [TXT] | Federation.html | 2021-06-15 16:20 | 59K | |
| [TXT] | GracefulDecommission.html | 2021-06-15 16:20 | 37K | |
| [TXT] | NodeAttributes.html | 2021-06-15 16:20 | 39K | |
| [TXT] | NodeLabel.html | 2021-06-15 16:20 | 45K | |
| [TXT] | NodeManager.html | 2021-06-15 16:20 | 50K | |
| [TXT] | NodeManagerCGroupsMemory.html | 2021-06-15 16:20 | 37K | |
| [TXT] | NodeManagerCgroups.html | 2021-06-15 16:20 | 34K | |
| [TXT] | NodeManagerRest.html | 2021-06-15 16:20 | 57K | |
| [TXT] | OpportunisticContainers.html | 2021-06-15 16:20 | 56K | |
| [TXT] | PlacementConstraints.html | 2021-06-15 16:20 | 48K | |
| [TXT] | PluggableDeviceFramework.html | 2021-06-15 16:20 | 34K | |
| [TXT] | ReservationSystem.html | 2021-06-15 16:20 | 31K | |
| [TXT] | ResourceManagerHA.html | 2021-06-15 16:20 | 41K | |
| [TXT] | ResourceManagerRest.html | 2021-06-15 16:20 | 345K | |
| [TXT] | ResourceManagerRestart.html | 2021-06-15 16:20 | 42K | |
| [TXT] | ResourceModel.html | 2021-06-15 16:20 | 42K | |

Scheduling settings and use resources

# Introduction

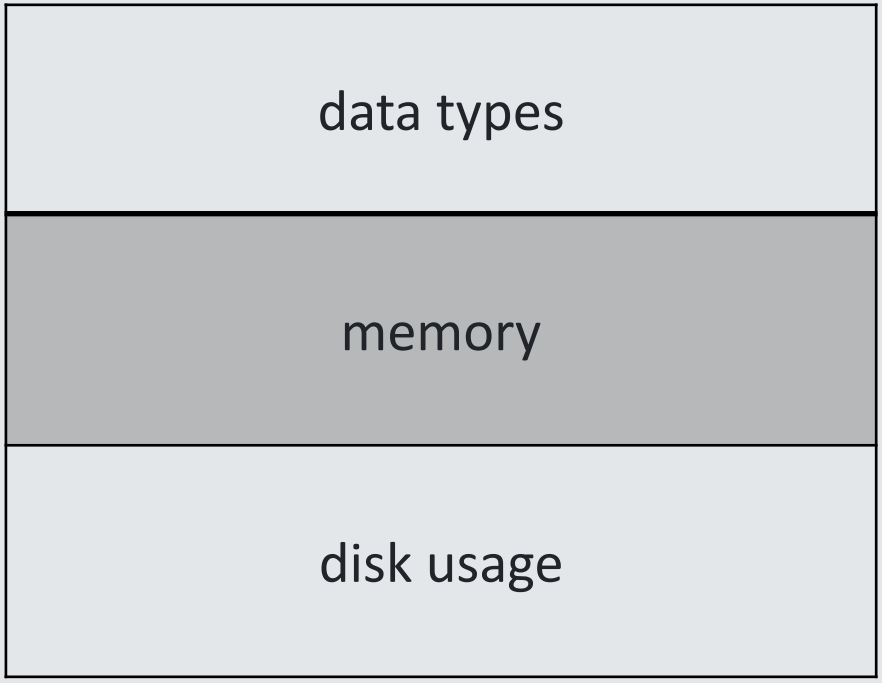
```
import java.io.IOException;
import java.util.StringTokenizer;

import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.Mapper;
import org.apache.hadoop.mapreduce.Reducer;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;

public class WordCount {

  public static class TokenizerMapper
       extends Mapper<Object, Text, Text, IntWritable>{

    private final static IntWritable one = new IntWritable(1);
    private Text word = new Text();

    public void map(Object key, Text value, Context context
                    ) throws IOException, InterruptedException {
      StringTokenizer itr = new StringTokenizer(value.toString());
      while (itr.hasMoreTokens()) {
        word.set(itr.nextToken());
        context.write(word, one);
      }
    }
  }
```

Java program for distributed computing on the cluster backend.

# Introduction

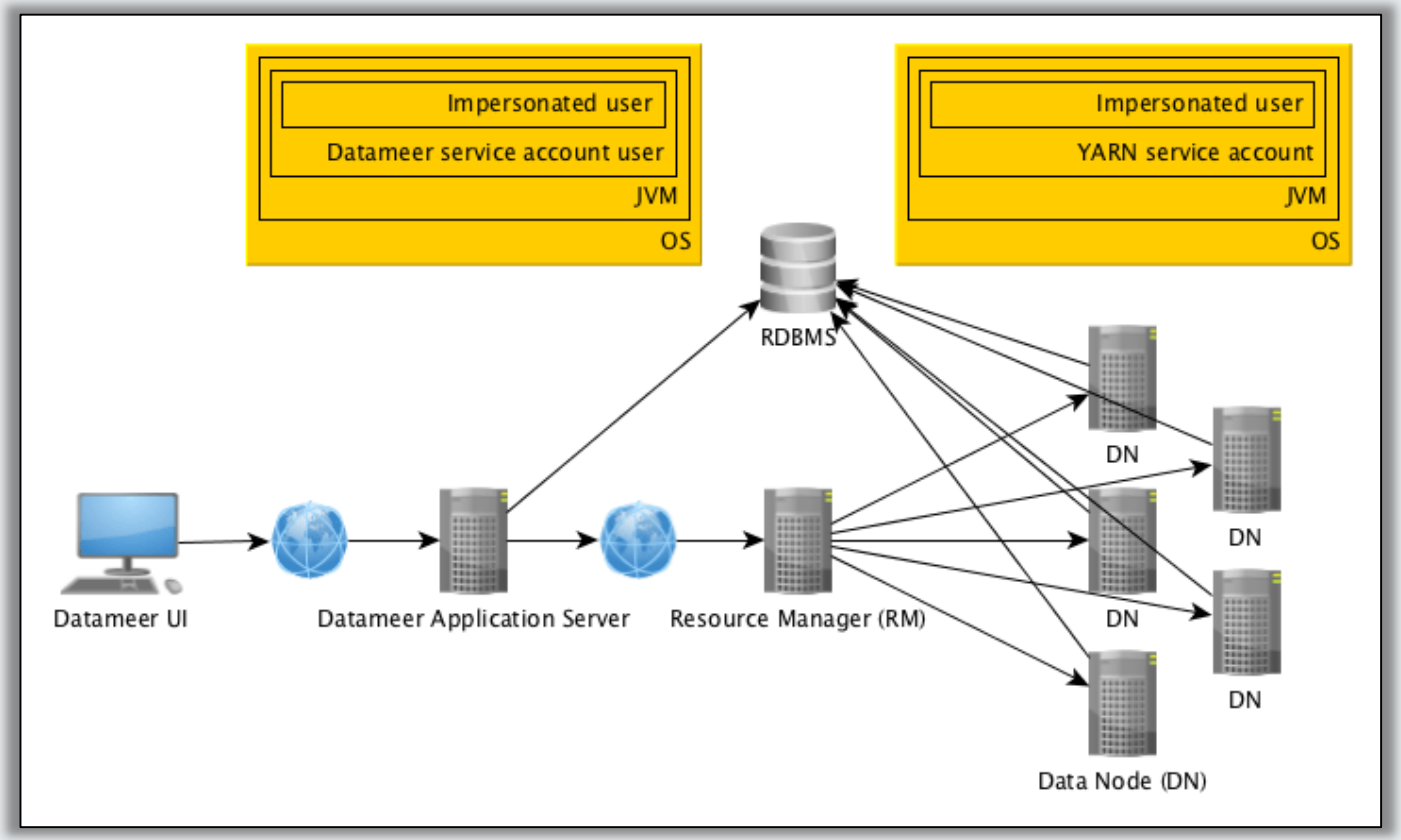

Datameer such an outstanding technology.

# Introduction



# Introduction

Analytics work into two stages




Design/edit time

Execution/runtime

# DESIGN/EDIT TIME

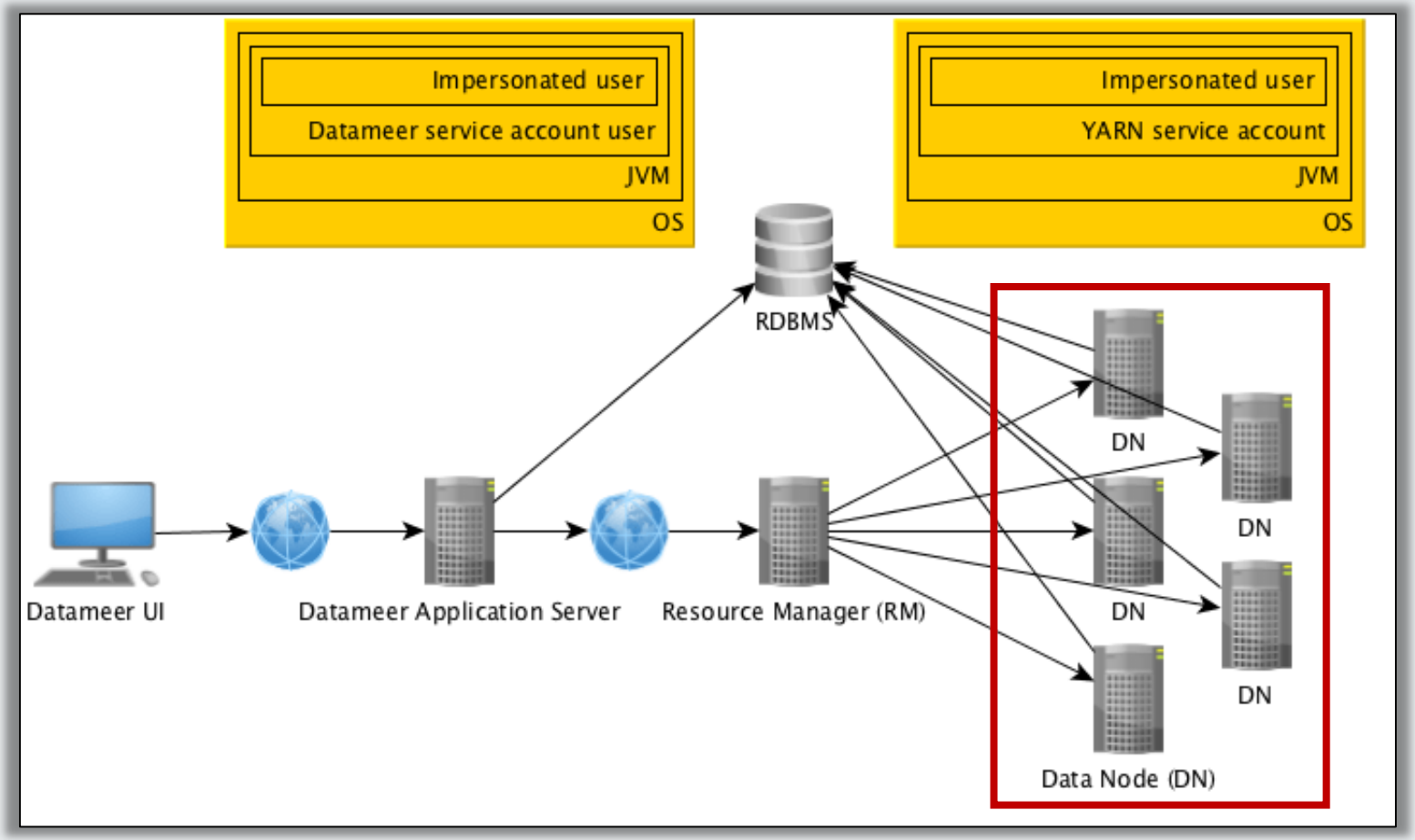

# DESIGN/EDIT TIME

<datameerServiceAccountUser>@<datameerHost>

or

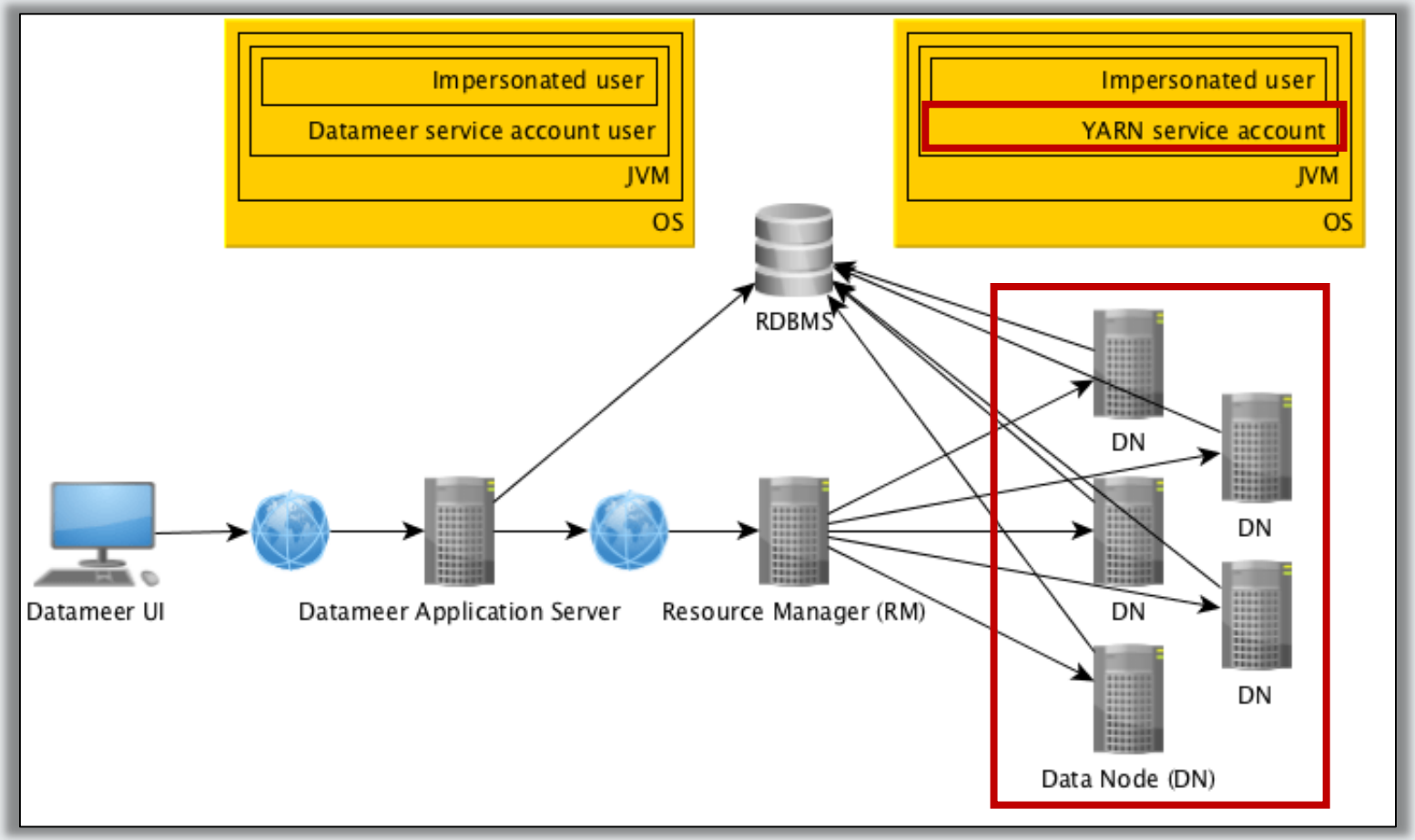
<loggedinUser>@<datameerHost>.

# EXECUTION/RUN TIME




Random DataNodes (DN) in the cluster

# EXECUTION/RUN TIME



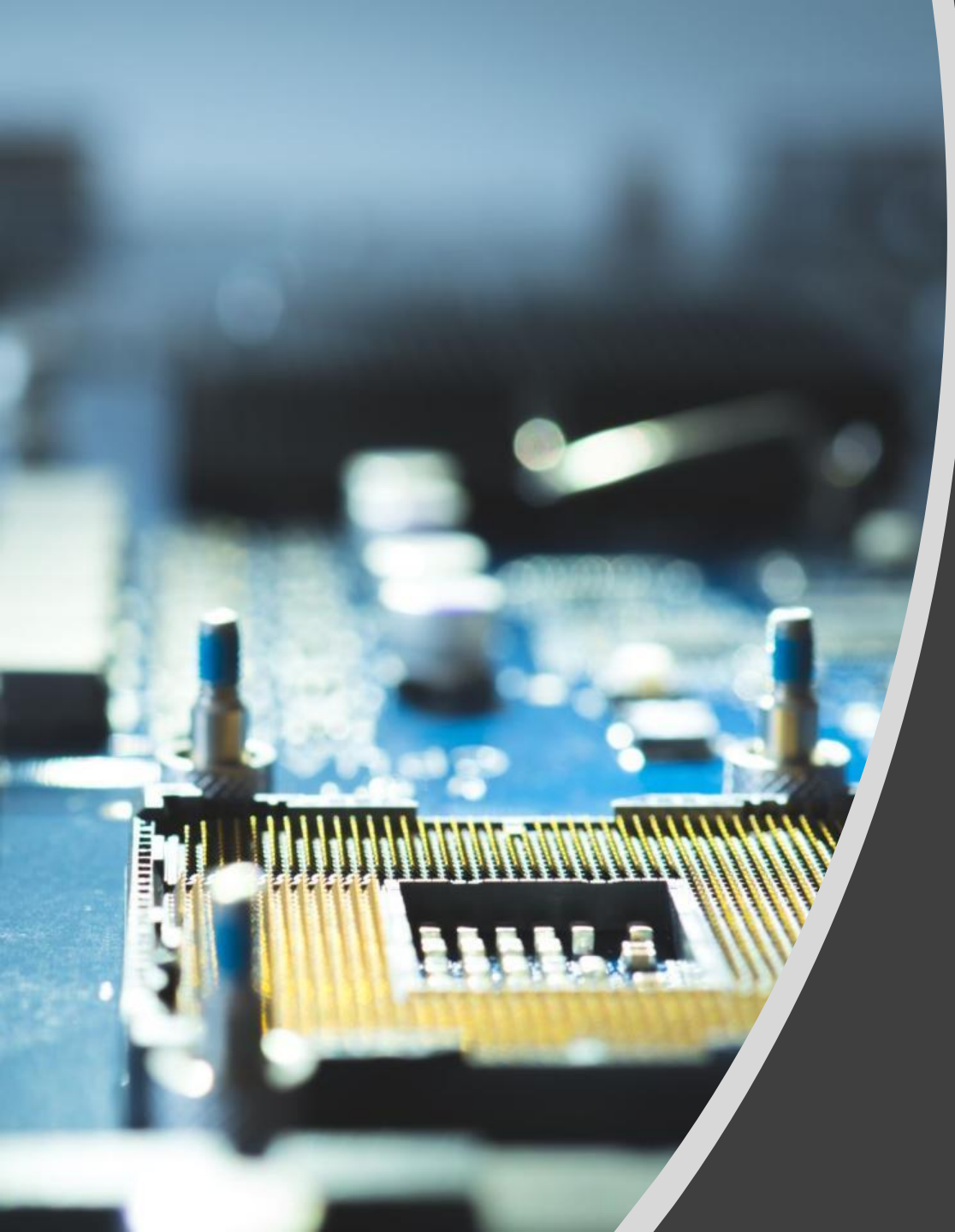

Random DataNodes (DN) in the cluster

# EXECUTION/RUN TIME

<yarnServiceAccountUser>@<dataNode>

or

<impersonatedUser>@<dataNode>.



# What is Data Preparation and Feature Engineering

cleaning,

structuring,

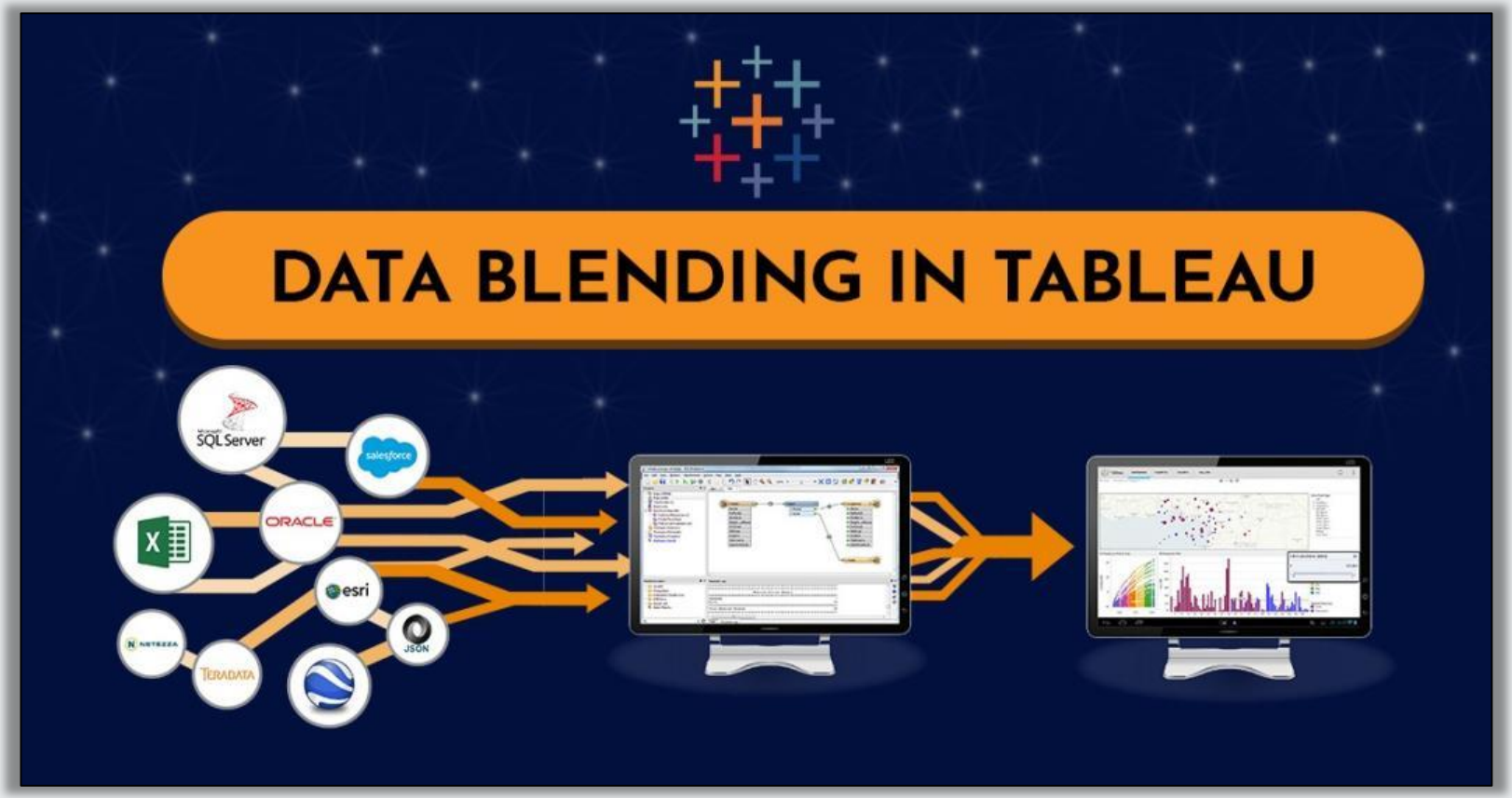
enriching raw data,

unstructured or big data.

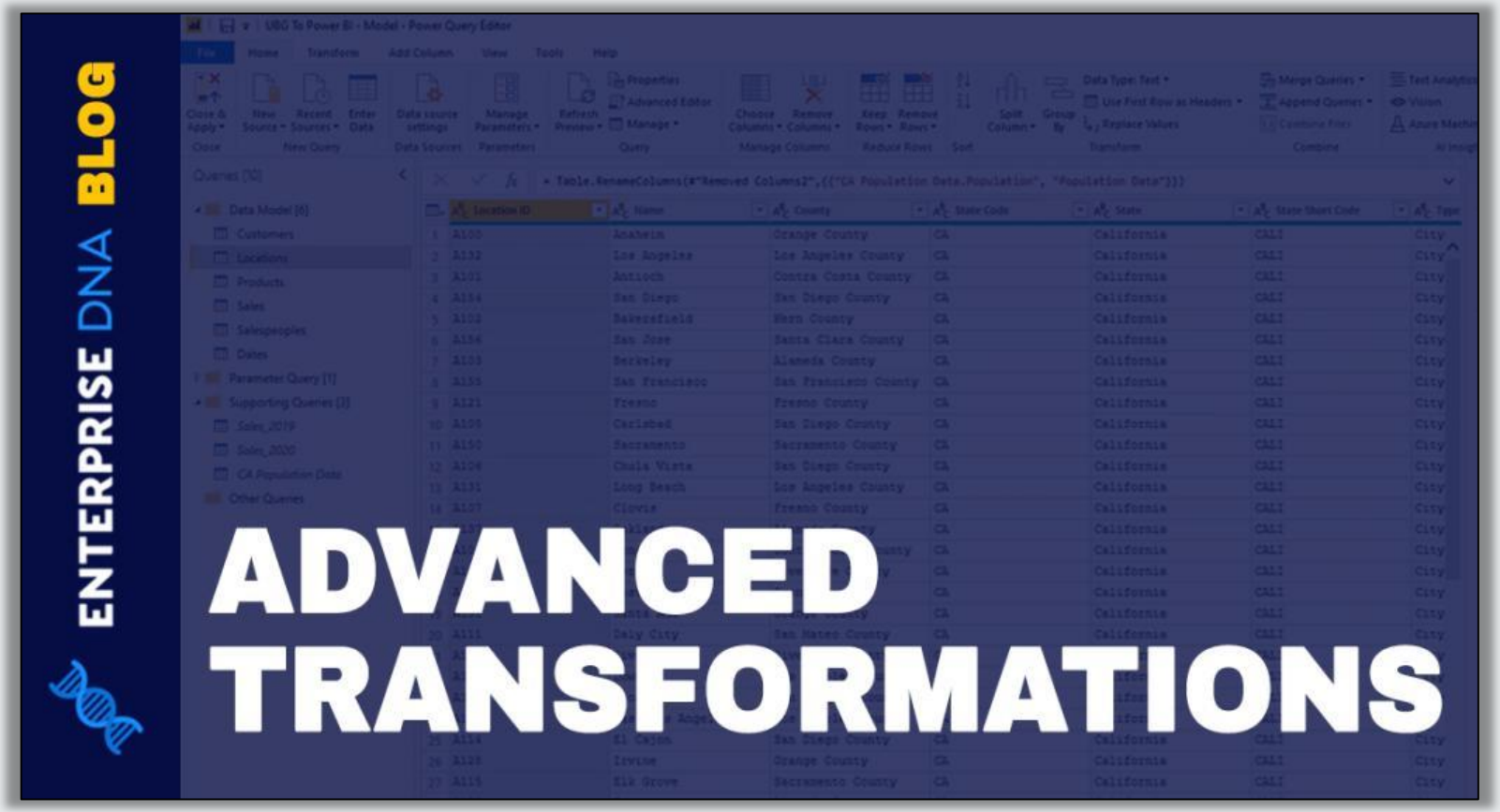
# Data Preparation in Datameer

# Data Cleansing

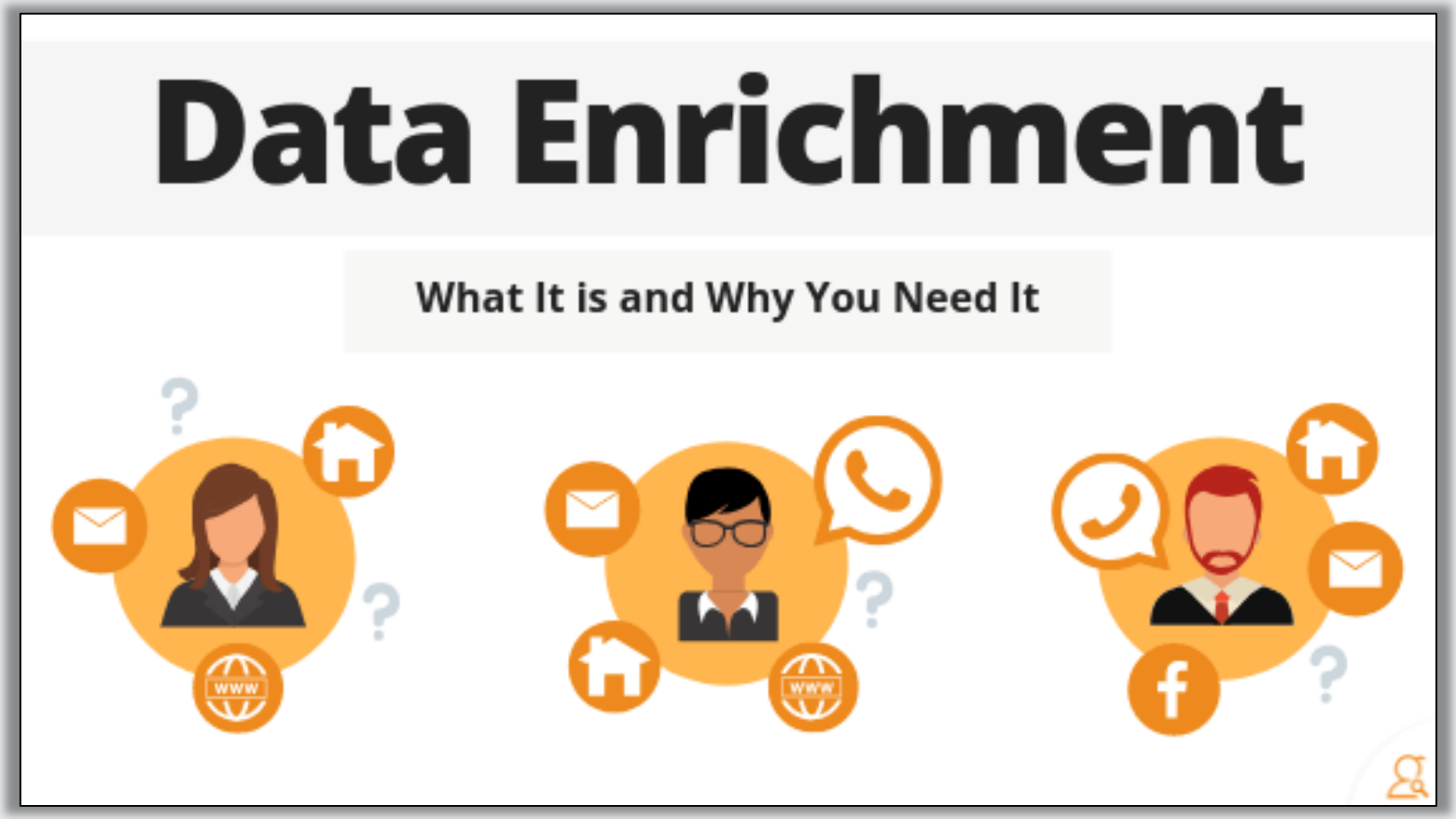
# Data Blending



# Data transformations



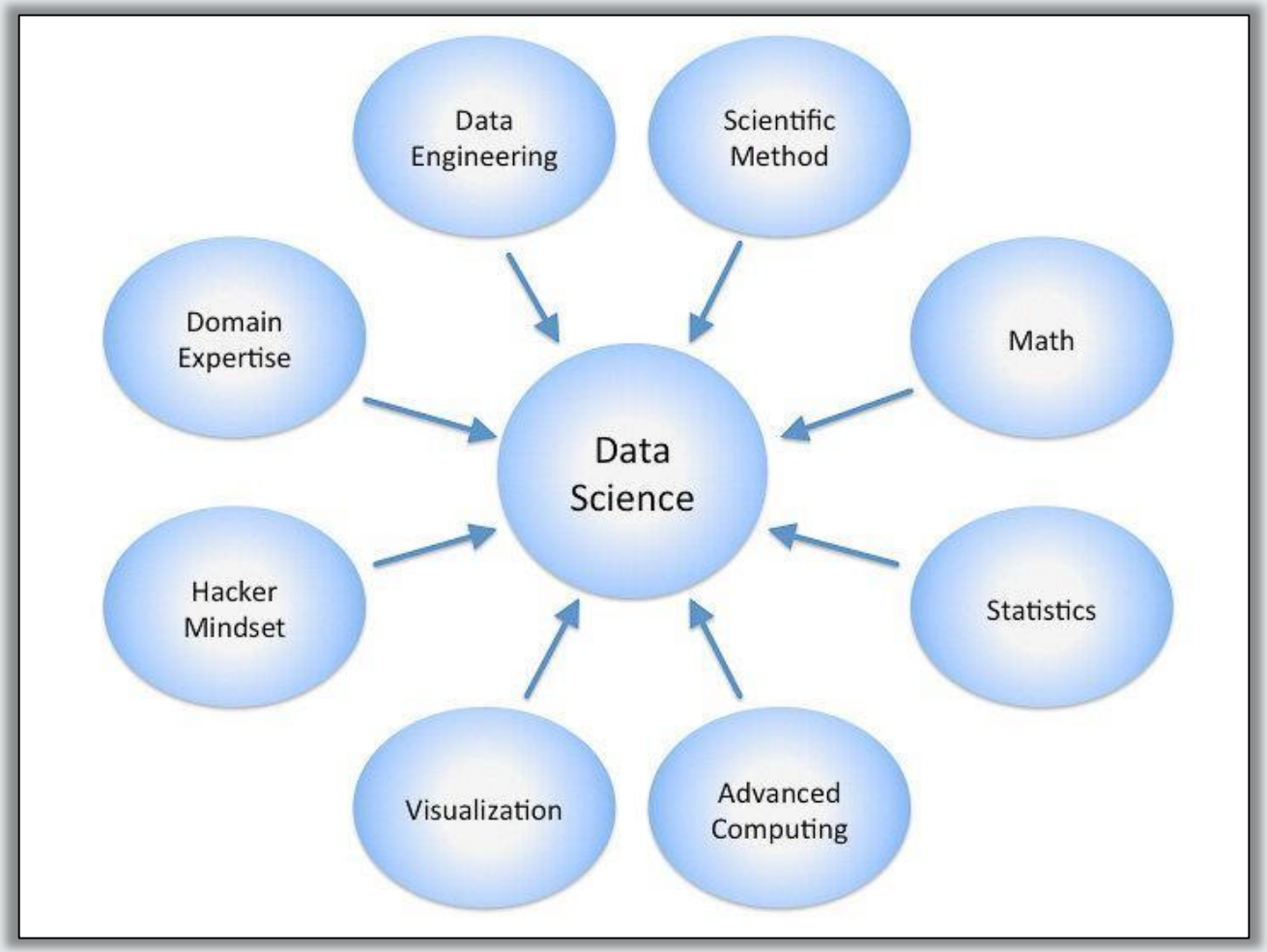
# Data enrichment



# Data grouping and organization

| Marks | Tally Marks | Frequency |
|---|---|---|
| 0 - 5 | 𝍸 I | 6 |
| 5 - 10 | 𝍸 𝍸 | 10 |
| 10 - 15 | 𝍸 III | 8 |
| 15 - 20 | 𝍸 IIII | 9 |
| 20 - 25 | 𝍸 II | 7 |
| | Total | 40 |

# Data science-specific




That's all for now...