

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 screenshot of Microsoft Excel demonstrating the use of the `FILTER` function. The formula in the formula bar is `=FILTER(A2:C13, B2:B13=F1, "No results")`. The table below shows the original data and the filtered results.

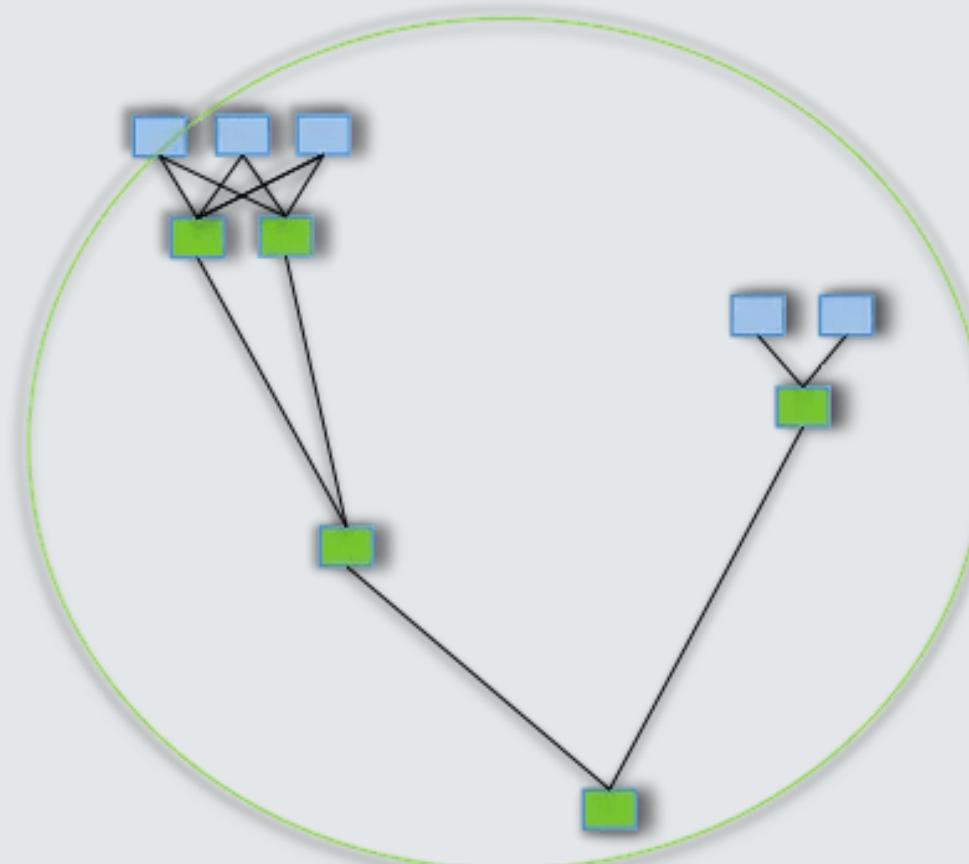
	A	B	C	D	E	F	G	H
1	Name	Group	Wins		Group	B		
2	Aiden	A	0					
3	Andrew	C	4					
4	Betty	B	1		Betty	B	1	
5	Caden	A	2		Charlotte	B	2	
6	Charlotte	B	2		Oliver	B	3	
7	Emma	C	0		Zoe	B	2	
8	Isabella	A	2					
9	Mason	A	4					
10	Nick	C	1					
11	Oliver	B	3					
12	Robert	C	3					
13	Zoe	B	2					

Introduction

	A	B	C	D	E	F	G	H
1	Name	Group	Wins		Group	B		
2	Aiden	A	0					
3	Andrew	C	4					
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					



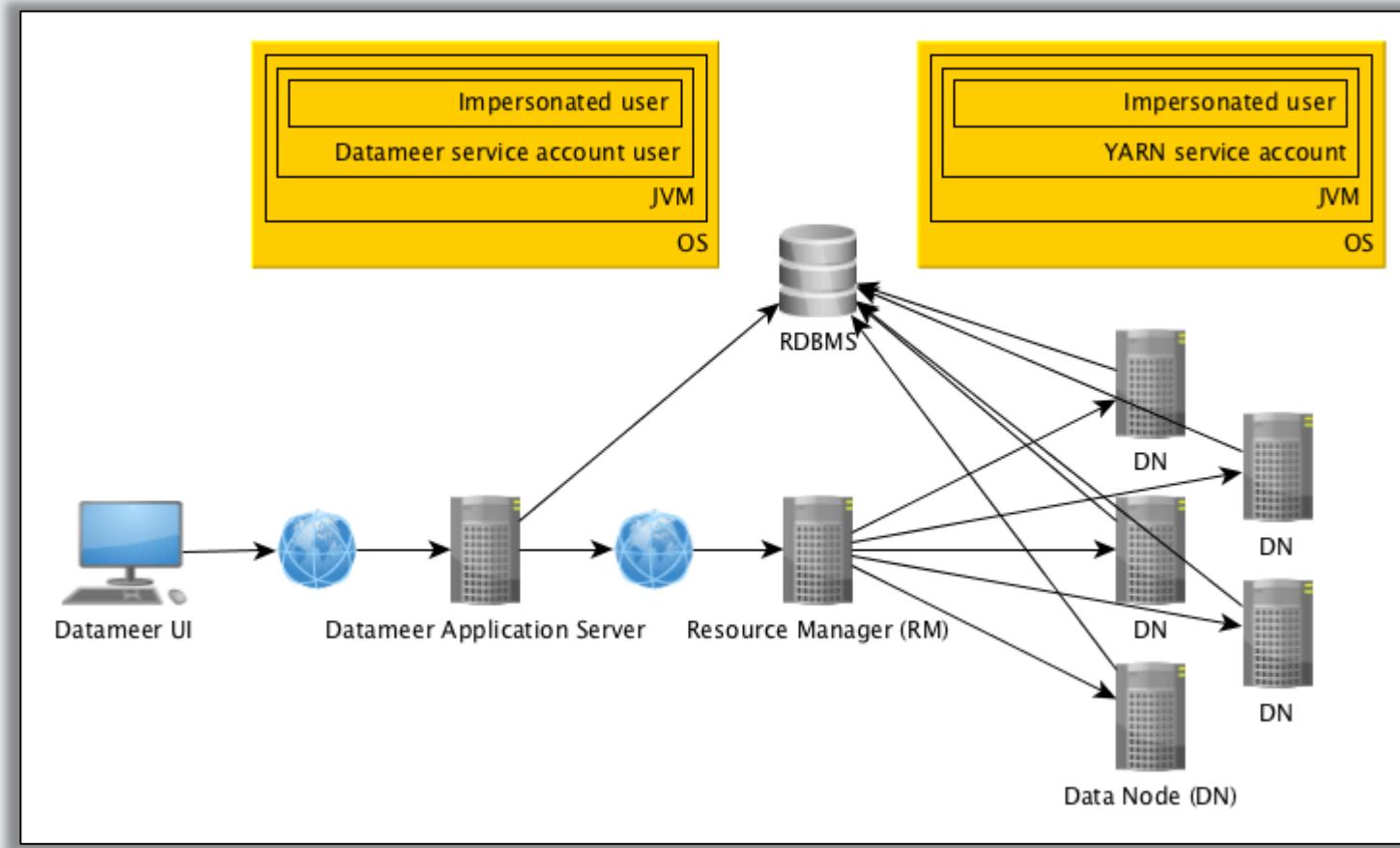
Introduction



Pig/Hive - Tez

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);
            }
        }
    }
}
```

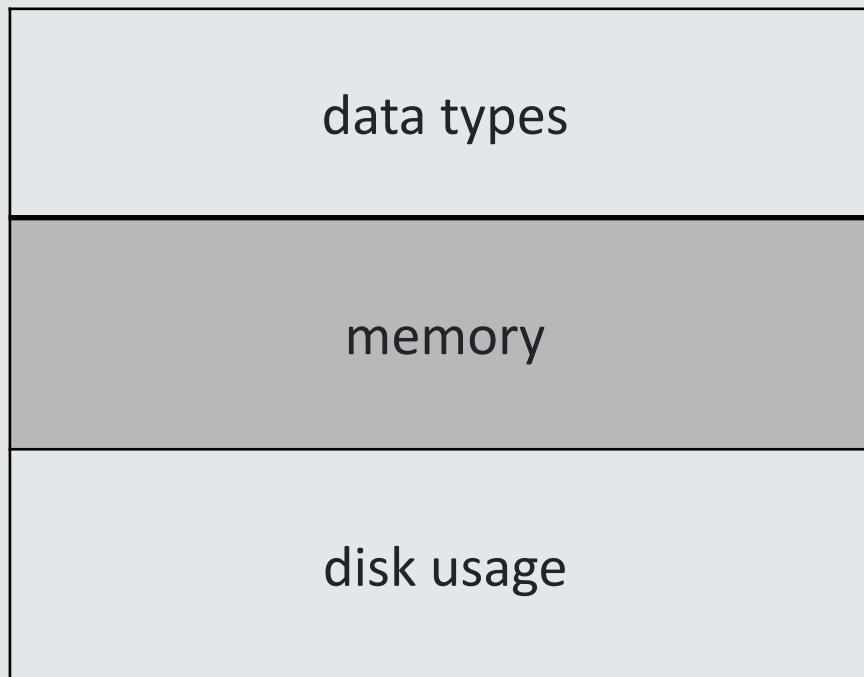
Java program for distributed computing on the cluster backend.

Introduction



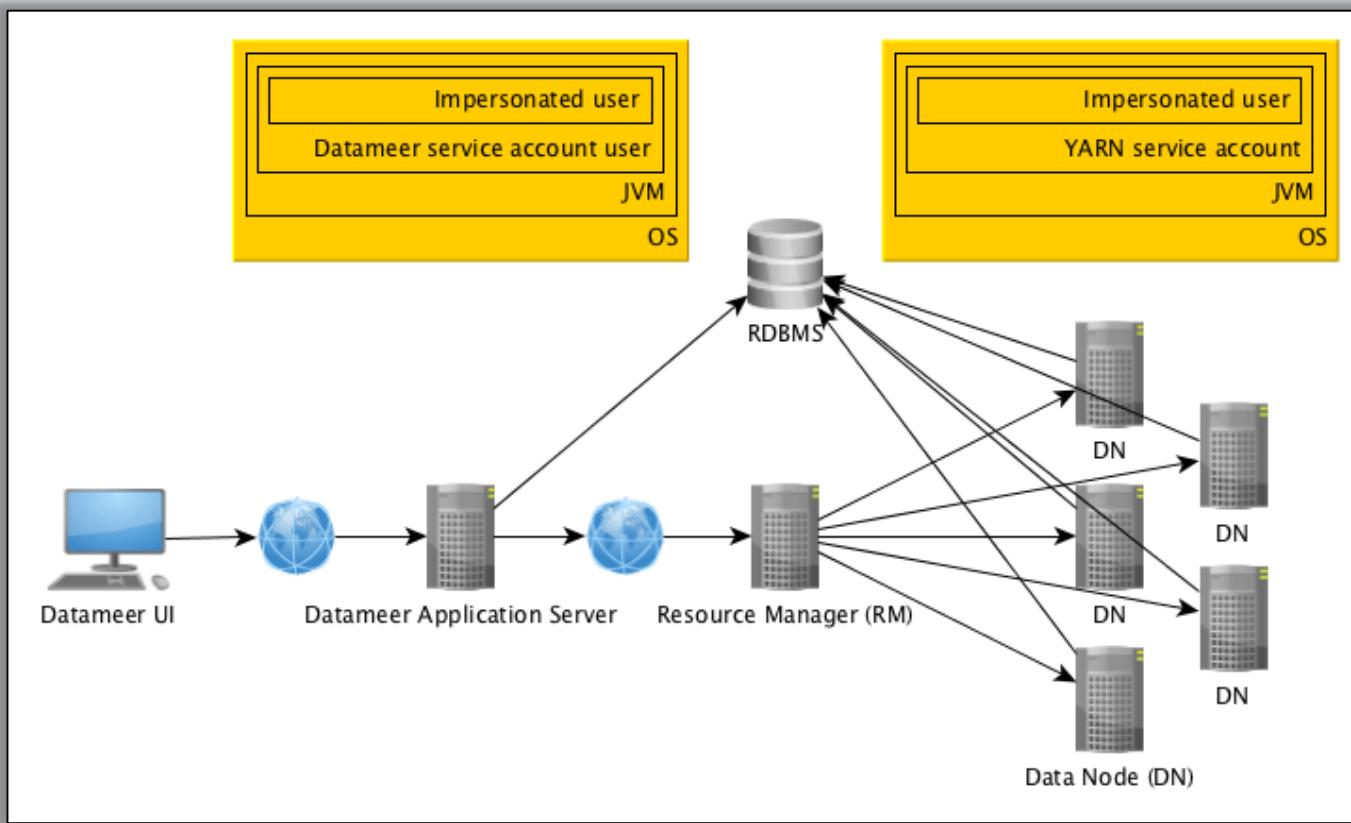
Datameer such an outstanding
technology.

Introduction



Introduction

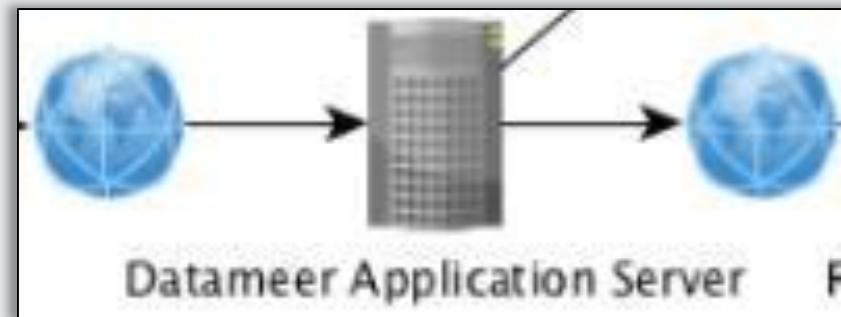
Analytics work into two stages



Design/edit time

Execution/runtime

DESIGN/EDIT TIME



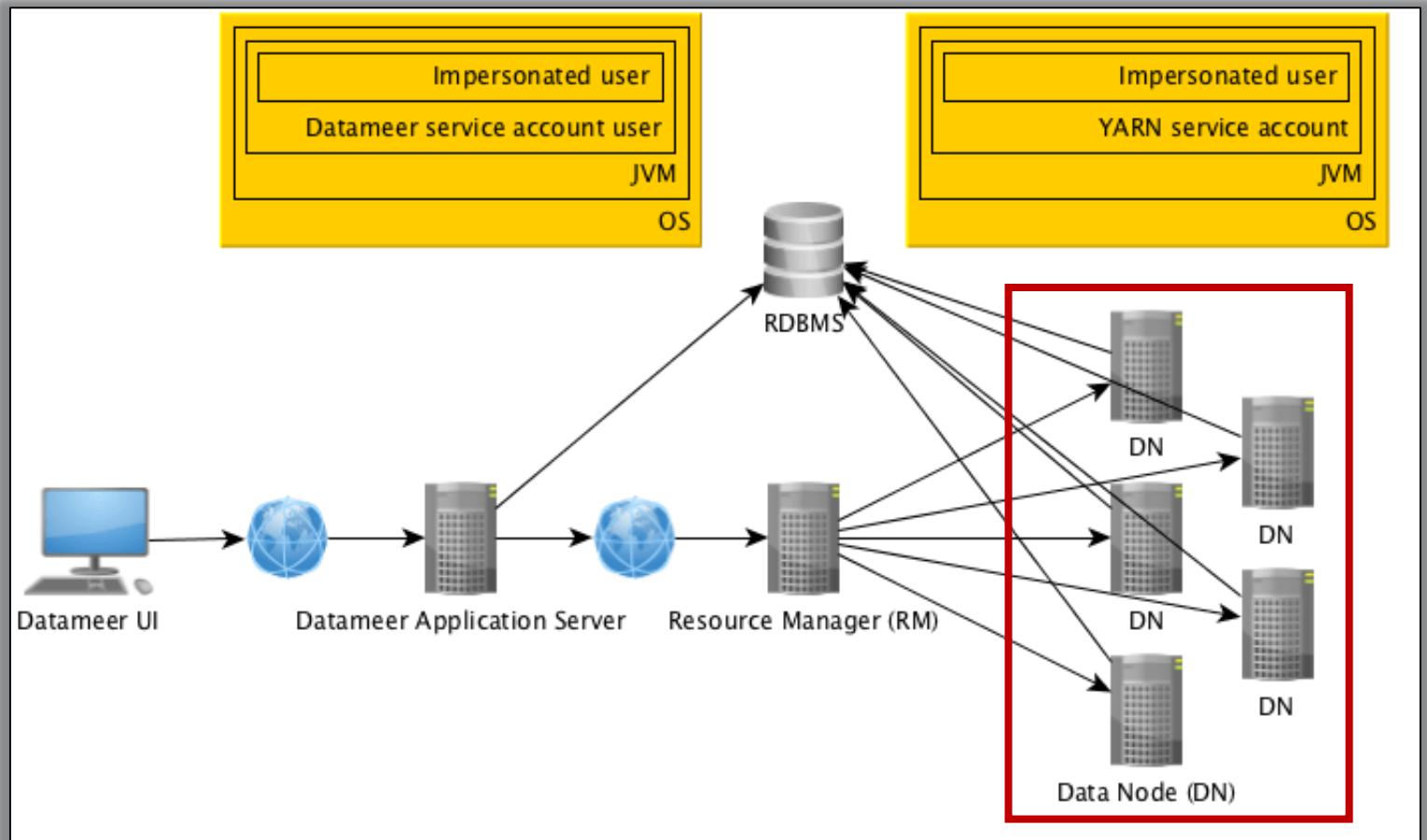
DESIGN/EDIT TIME

<datameerServiceAccountUser>@<datameerHost>

or

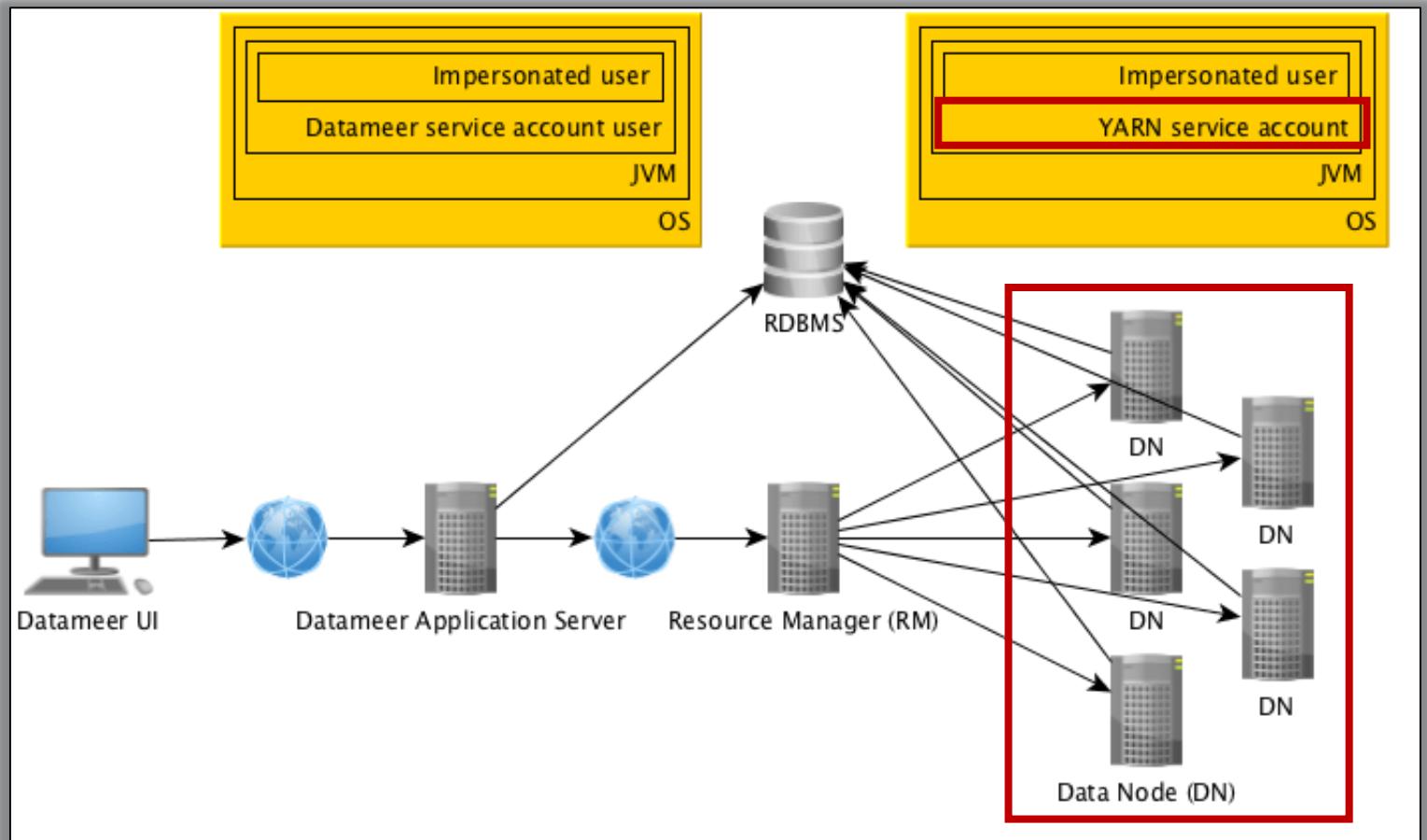
<loggedinUser>@<datameerHost>.

EXECUTION/RUN TIME



Random DataNodes (DN) in the cluster

EXECUTION/RUN TIME



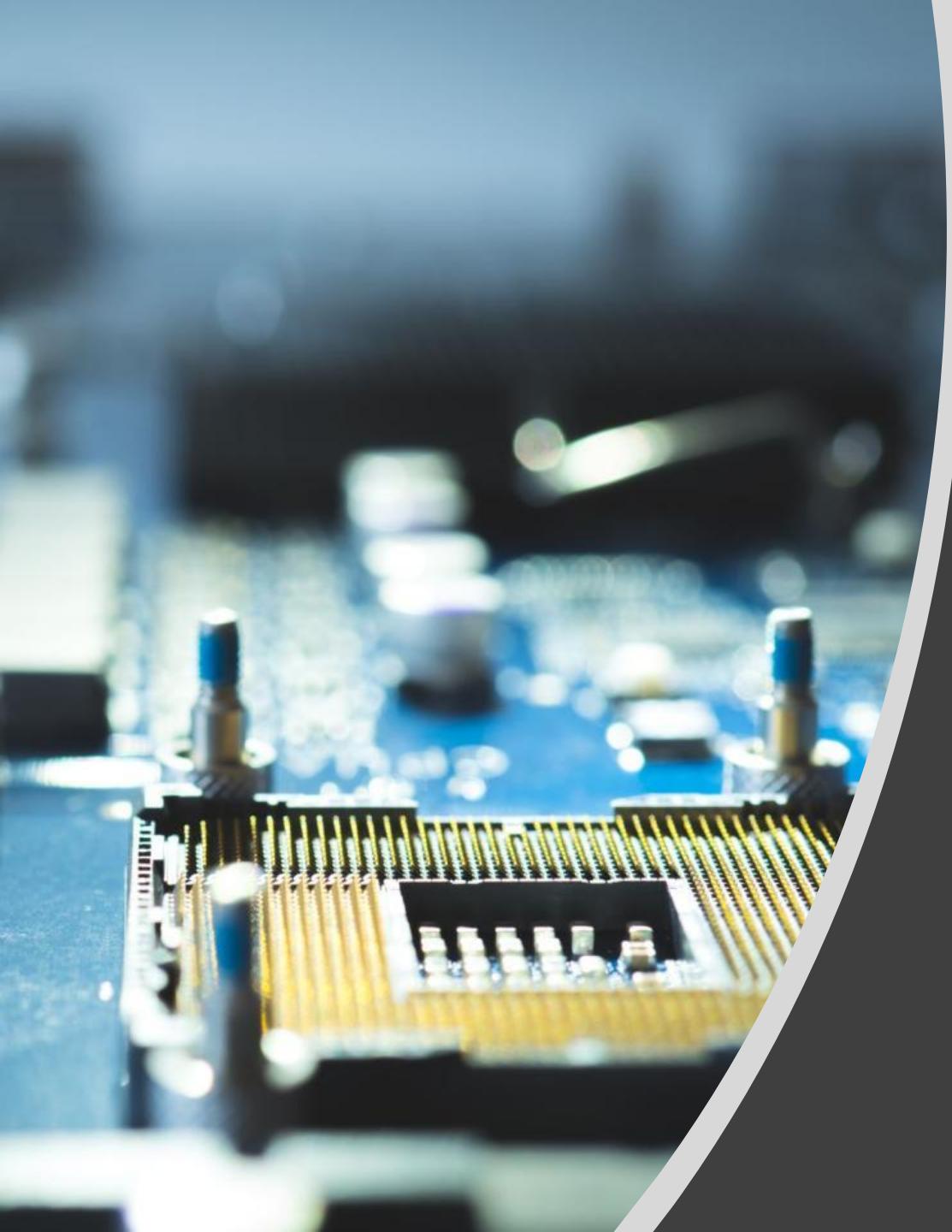
Random DataNodes (DN) in the cluster

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

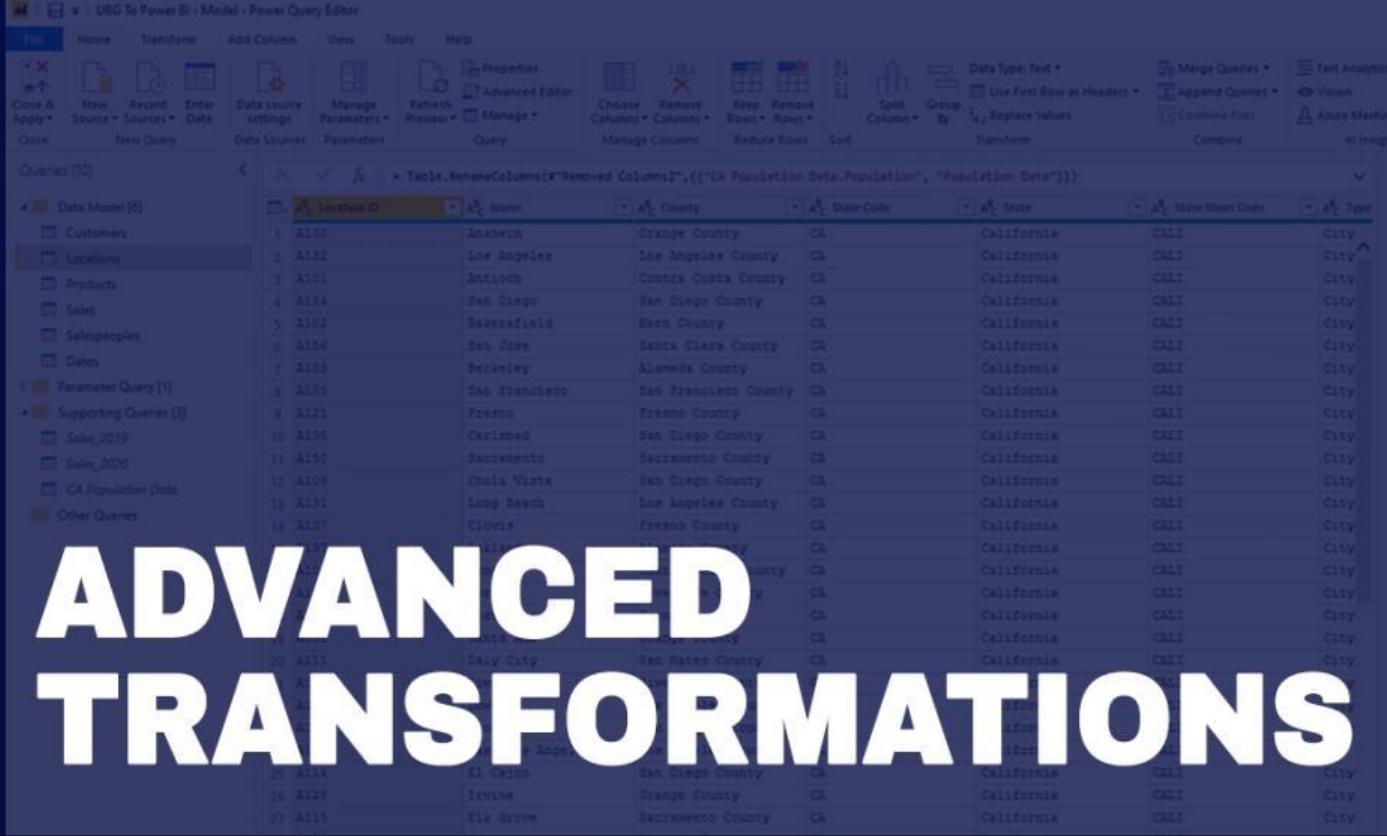


Data Blending



Data transformations

ENTERPRISE DNA BLOG



The screenshot shows the Microsoft Power Query Editor interface. On the left, the 'Queries' pane lists several queries: Data Model [6] (Customers, Locations, Products, Sales, Salespeople, Dates), Parameter Query [1], Supporting Queries [3] (Sales_2019, Sales_2020, CA Population Data), and Other Queries. The main area displays a table titled 'Table.RenameColumns("Removed Columns2", {"CA Population Data.Population", "Population Data"})'. The table contains 27 rows of data about California locations, with columns: LocationID, Name, County, StateCode, State, StateShortCode, and Type. The data includes rows for various cities like Anaheim, Los Angeles, Antioch, San Diego, Bakersfield, San Jose, Berkeley, San Francisco, Fresno, Sacramento, Chula Vista, Long Beach, Glendale, Redlands, Ontario, Pomona, San Bernardino, and El Cajon. The 'Type' column consistently shows 'City'. The Power Query ribbon at the top has tabs for File, Home, Transform, Add Column, View, Tools, and Help, along with various transformation tools like Refresh, Advanced Editor, Manage, and Split Column.

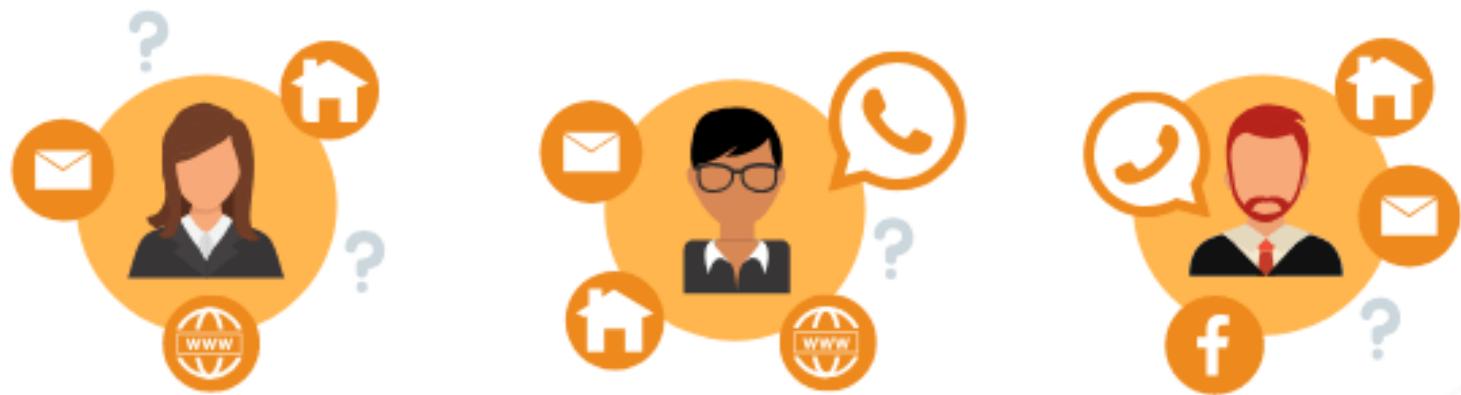
ADVANCED
TRANSFORMATIONS

LocationID	Name	County	StateCode	State	StateShortCode	Type
A100	Anaheim	Orange County	CA	California	CA	City
A102	Los Angeles	Los Angeles County	CA	California	CA	City
A103	Antioch	Contra Costa County	CA	California	CA	City
A104	San Diego	San Diego County	CA	California	CA	City
A105	Bakersfield	Kern County	CA	California	CA	City
A106	San Jose	Santa Clara County	CA	California	CA	City
A107	Berkeley	Alameda County	CA	California	CA	City
A108	San Francisco	San Francisco County	CA	California	CA	City
A109	Fresno	Fresno County	CA	California	CA	City
A110	Carlsbad	San Diego County	CA	California	CA	City
A111	Sacramento	Sacramento County	CA	California	CA	City
A112	Chula Vista	San Diego County	CA	California	CA	City
A113	Long Beach	Los Angeles County	CA	California	CA	City
A114	Glendale	Fresno County	CA	California	CA	City
A115	Redlands	Riverside County	CA	California	CA	City
A116	Ontario	San Bernardino County	CA	California	CA	City
A117	Pomona	San Bernardino County	CA	California	CA	City
A118	San Bernardino	San Bernardino County	CA	California	CA	City
A119	El Cajon	San Diego County	CA	California	CA	City
A120	Irvine	Orange County	CA	California	CA	City
A121	Ela Grove	Sacramento County	CA	California	CA	City

Data enrichment

Data Enrichment

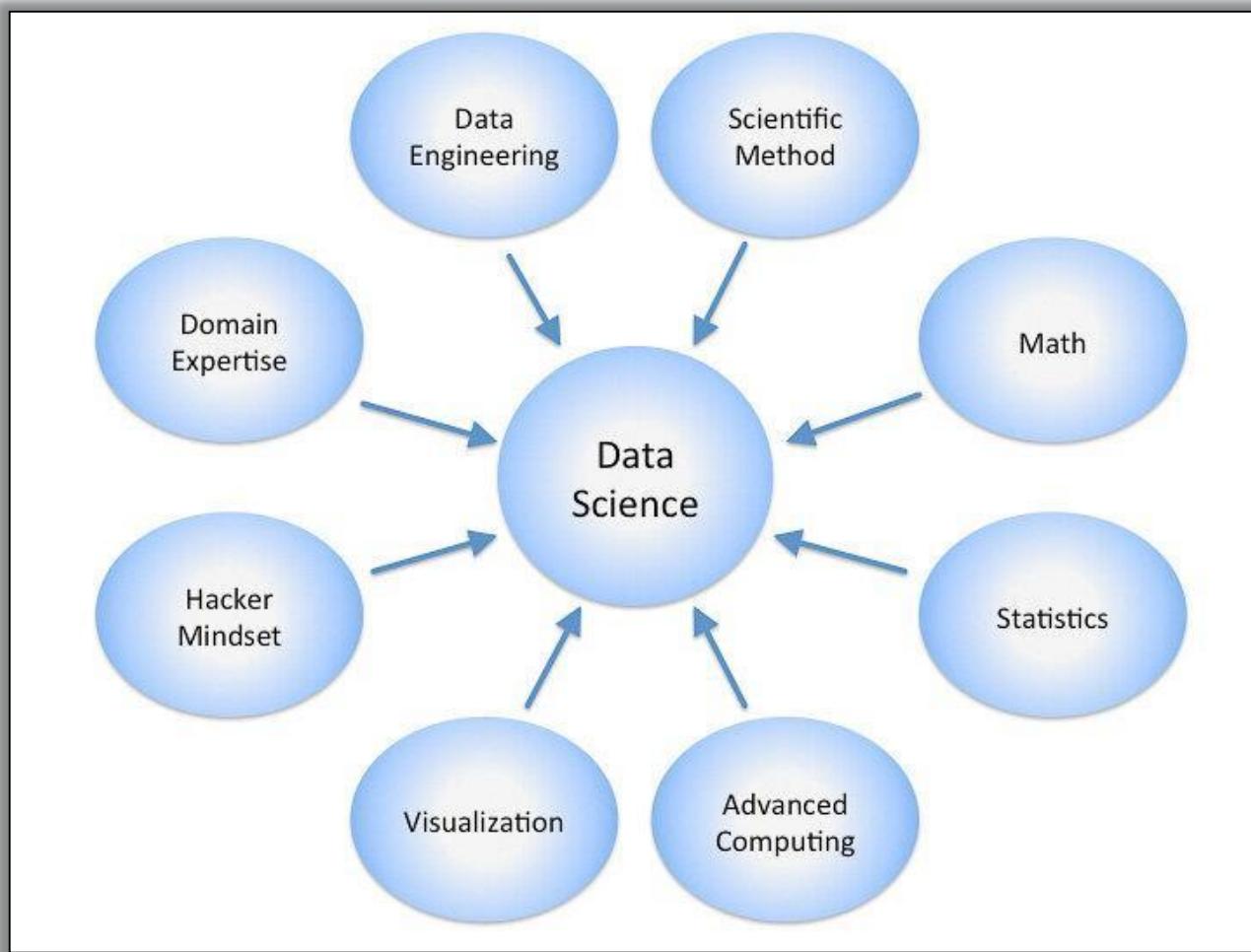
What It is and Why You Need It



Data grouping and organization

Marks	Tally Marks	Frequency
0 - 5		6
5 - 10		10
10 - 15		8
15 - 20		9
20 - 25		7
Total		40

Data science-specific





That's all for now...