CIT650: Introduction to Big Data

Lab #3

Lab Goals

- Spark on Docker
- Spark RDDs
- Spark DataFrames
- Spark SQL
- Spark MLlib

Spark on Docker using

- Pull the following image: docker pull jupyter/all-spark-notebook
- Create a container as follows:

```
docker container run -dit --name sparkj -p 8889:8889 -p 4040:4040 -p 4041:4041 jupyter/all-spark-notebook

docker exec -it sparkj bash
```

- Create a jupyter notebook inside the container and start play with Spark!

Start using Spark

- *Let's open the lab's notebook.*

Thanks