# —— Oozie, Pig, MapReduce and HCatalog —— Computational Infrastructure for Big Data Analytics

L. Felipe Rivera

School of Engineering

July, 2017

# Outline I

1. Context
   - Project objectives

# Project objectives

## What are we looking for?

| Question | Condition | Related QAs | Ecosystem Apps | Dataset | Program(s) | Current State |
|---|---|---|---|---|---|---|
| How big is the gap between Pig and MapReduce execution times? | Proposed | Performance | Pig, MapReduce, Oozie | NCDC Weather | Max temperature and Mean maximun temperature station-day-month | Executed on cluster |
| Can Pig exploit the benefits[1] of Hive through HCatalog? | Proposed | Performance, Extensibility | Pig, HCatalog | NCDC Weather | Partitioned weather | Tested on Cluster |
| How easily does Oozie support changes in workflow apps? | Proposed | Reusability, Maintainability | Pig, MapReduce, Oozie | NCDC Weather | Mean maximun temperature station-day-month | Tested on Cloudera VM |
| What is the gap between Oozie and JobControl execution times? | Desired | Performance | Oozie, MapReduce | NCDC Weather | Mean maximun temperature station-day-month | JobControl instance to be coded, tested and executed. |

[1] Benefits provided by partitioned or bucketed tables.

# Bibliography I