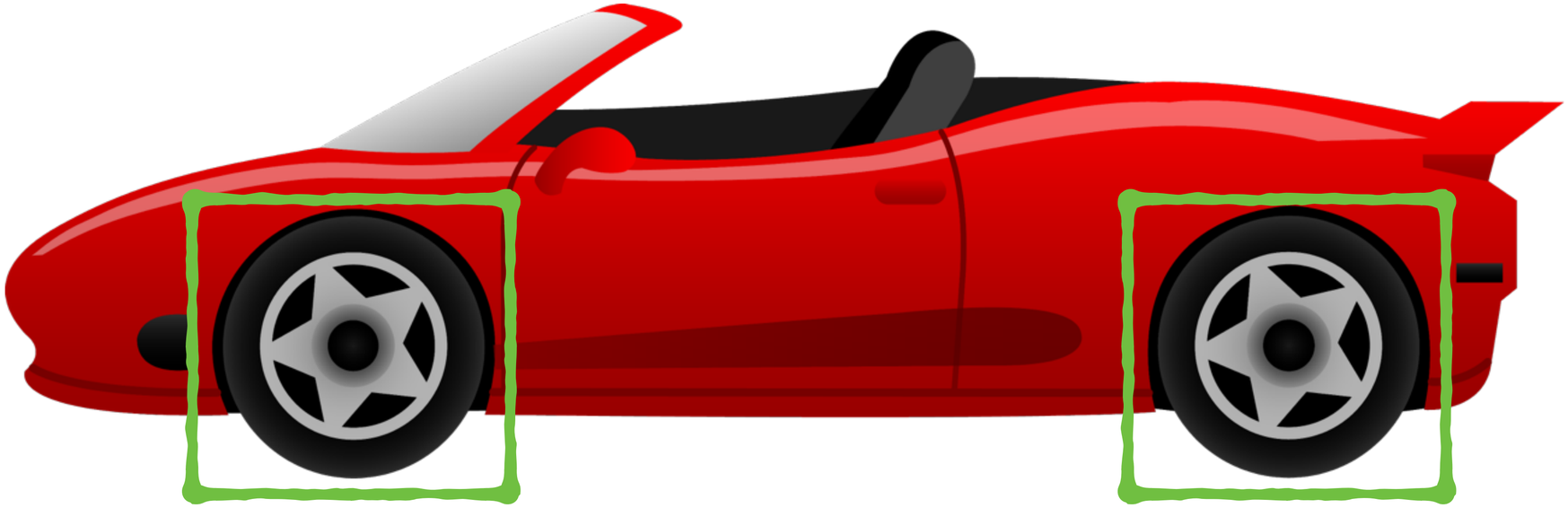
OOZIE

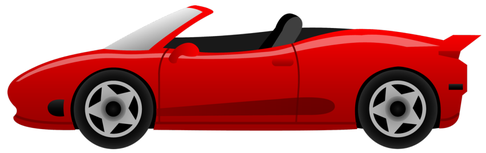
OOZIE

Oozie is a cute name but what does it do?

OOZIE

Consider a factory which manufactures wheels for cars

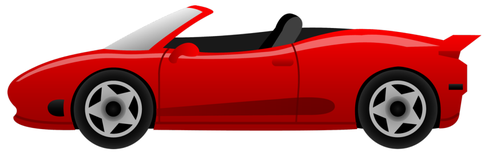




OOZIE

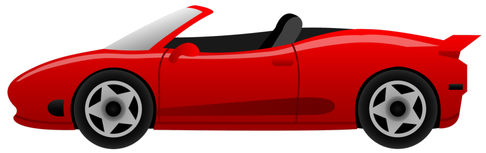
It manufactures the tires

It manufactures the hubcaps



OOZIE

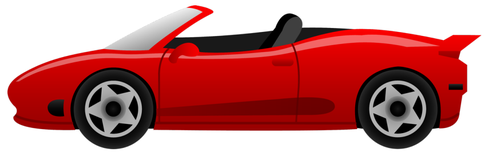
It procures tools



OOZIE

It puts them together

To get a wheel!



OOZIE

We may have slightly simplified the entire process

But you get the idea

OOZIE

Every end is accomplished by completing a series of tasks

Some tasks may be done **serially**

Some tasks may be done **in parallel**

serially

in parallel

OOZIE

Some tasks may be **dependent** on other tasks

Some tasks may be **independent**

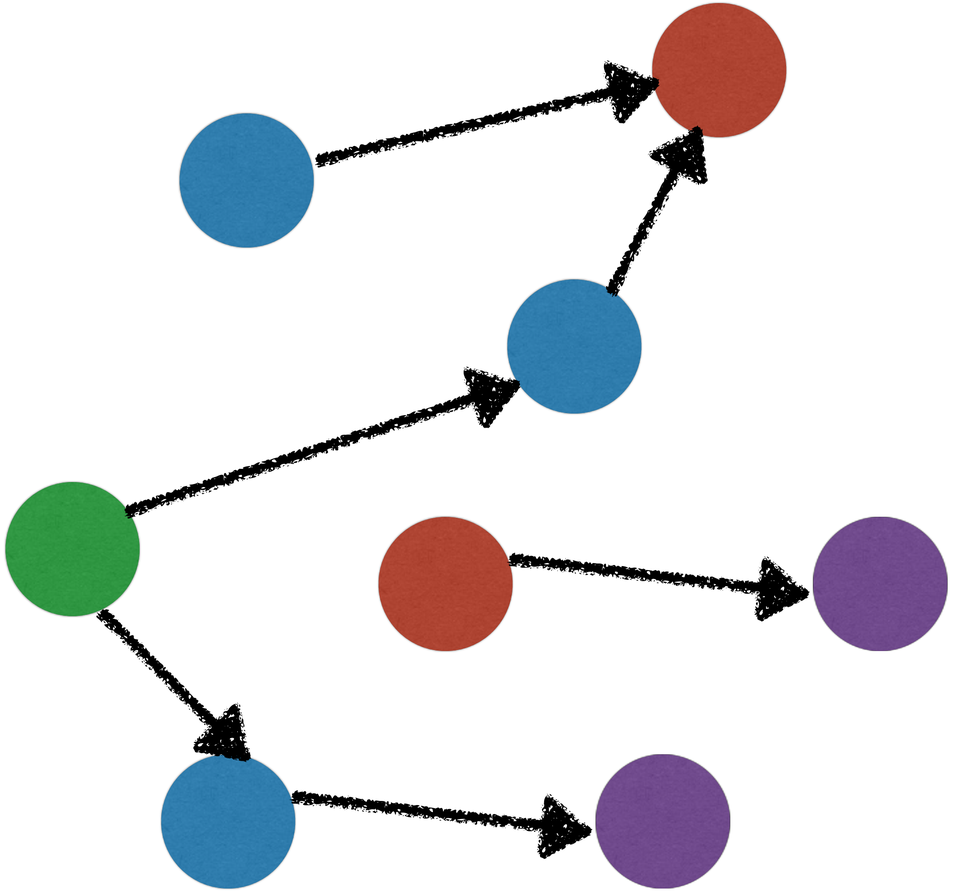
OOZIE

serially

in parallel

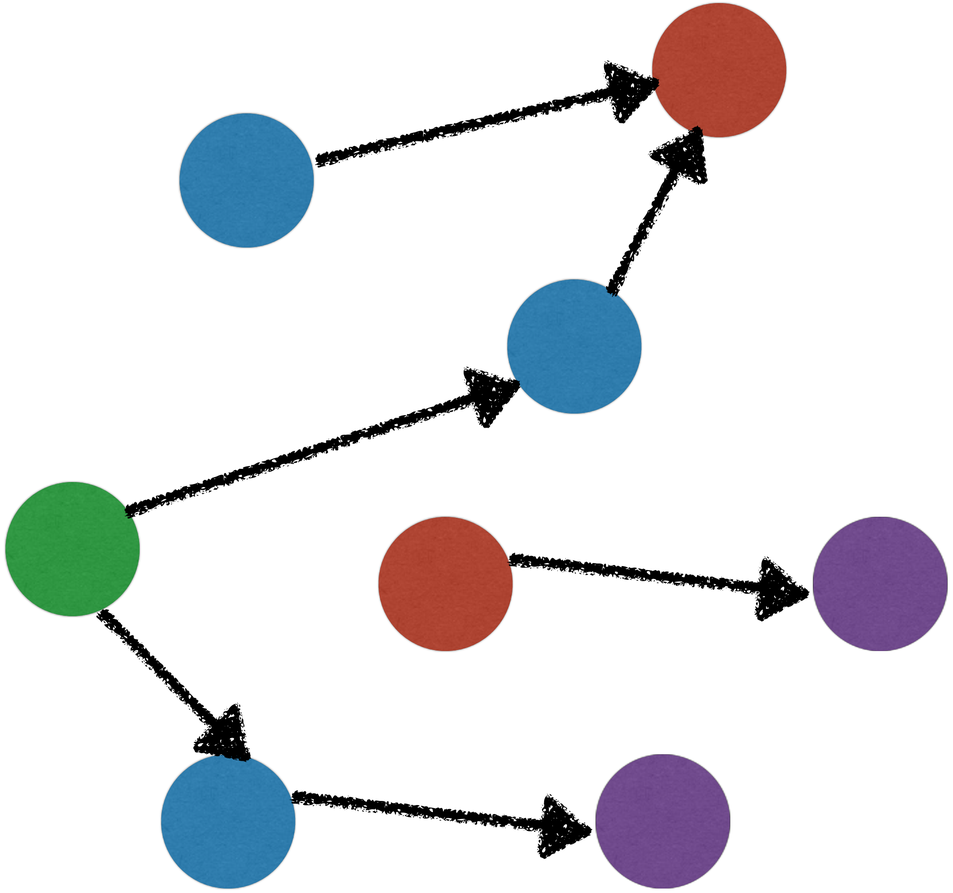
dependent

independent



OOZIE

Each node here represents a task



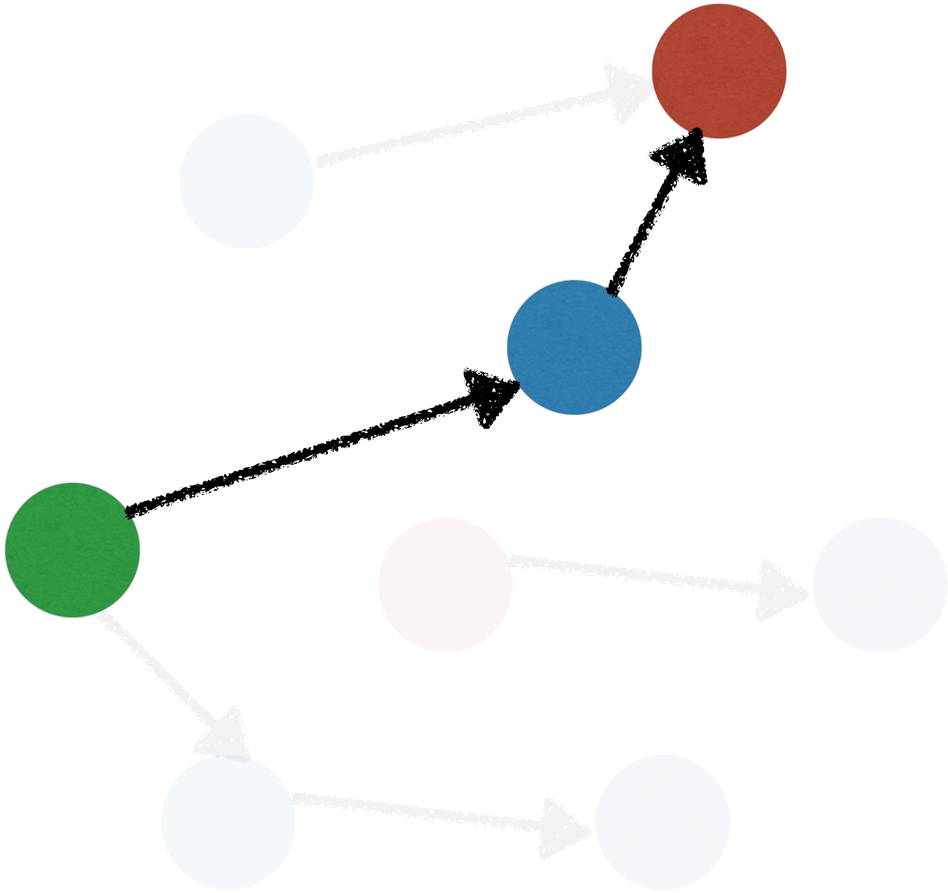
OOZIE

serially

in parallel

dependent

independent



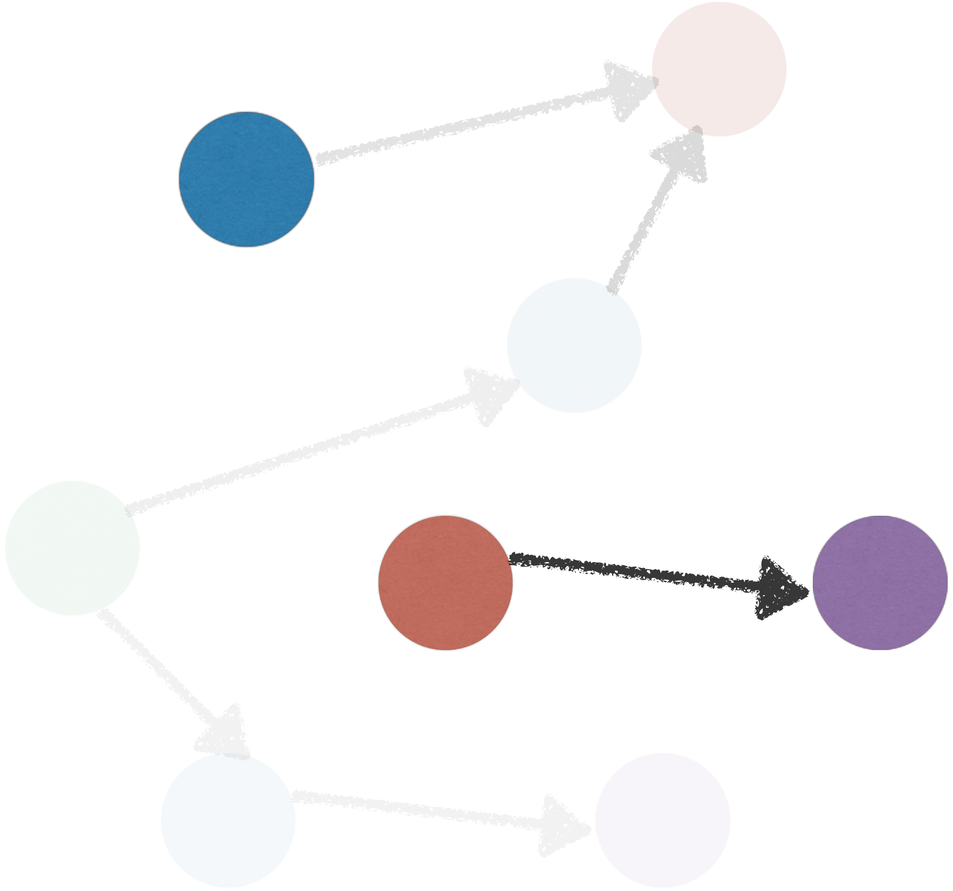
OOZIE

serially

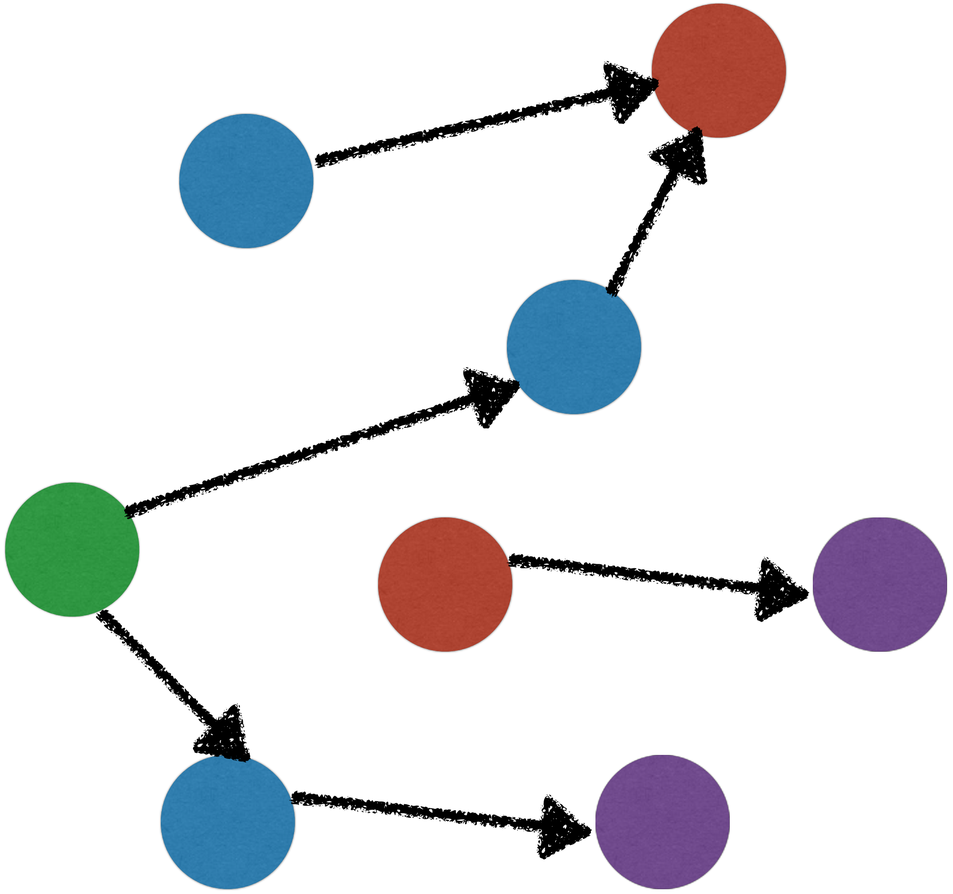
in parallel

dependent

independent

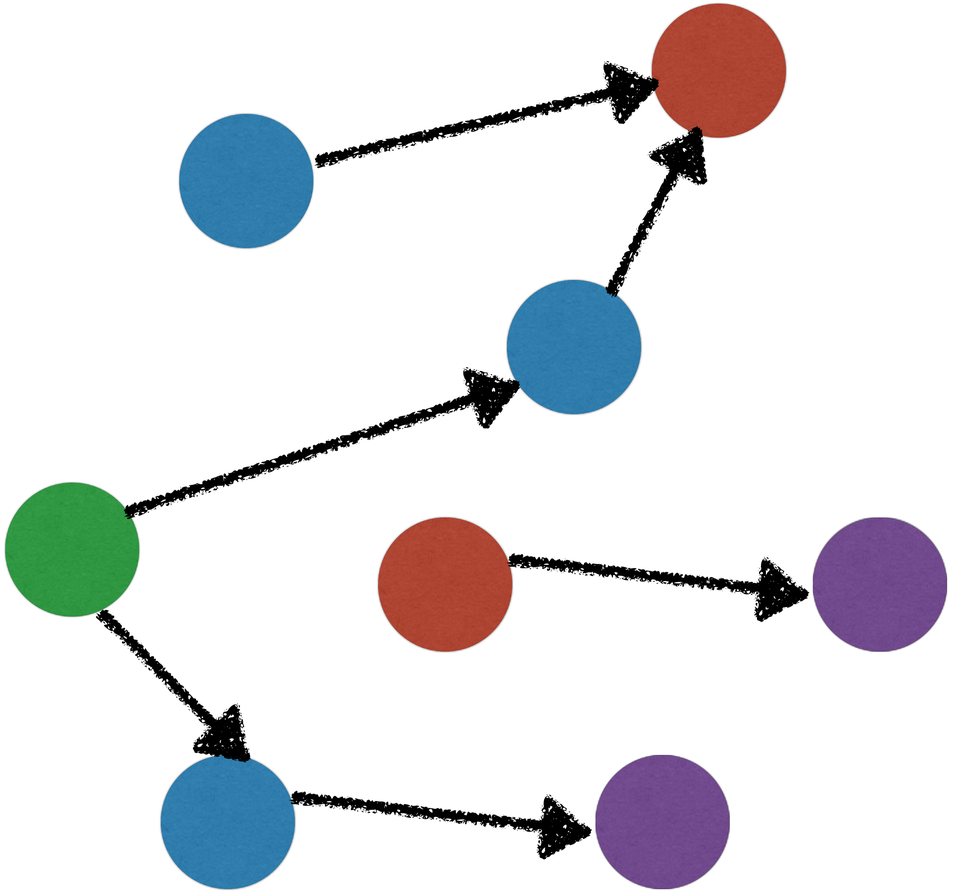


OOZIE

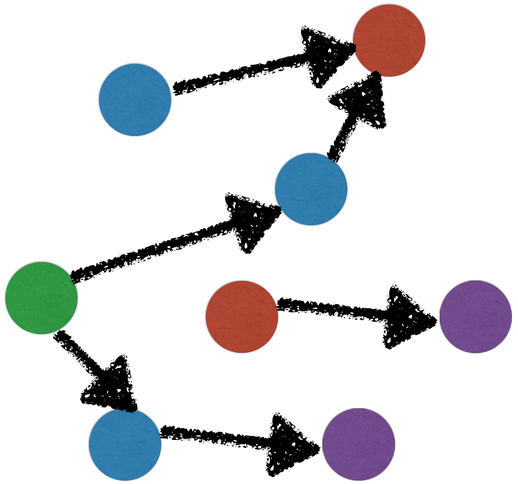


Directed Acyclic Graph

OOZIE



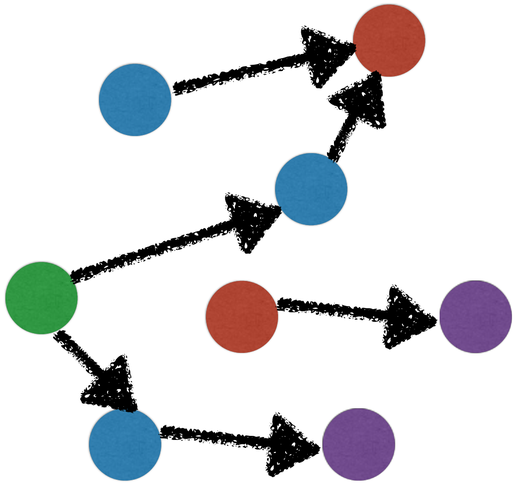
This is a Workflow



OOZIE

This is a Workflow

A workflow specifies a set of actions and the order and conditions under which those actions should be performed



OOZIE

This is a Workflow

A workflow specifies a set of actions and the order and conditions under which those actions should be performed

Hive Query

MapReduce Job

Pig Query

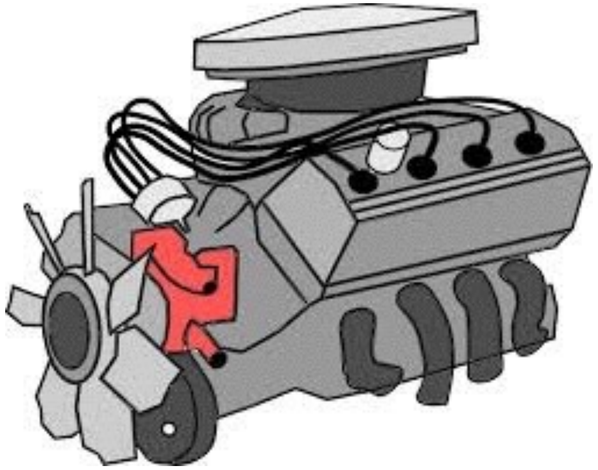
Shell Scripts

Java Program

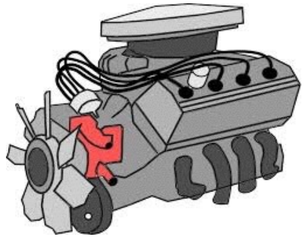
OOZIE

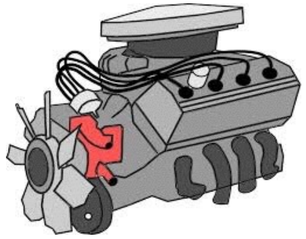
Now, if you're manufacturing cars, getting tires alone is not enough

OOZIE



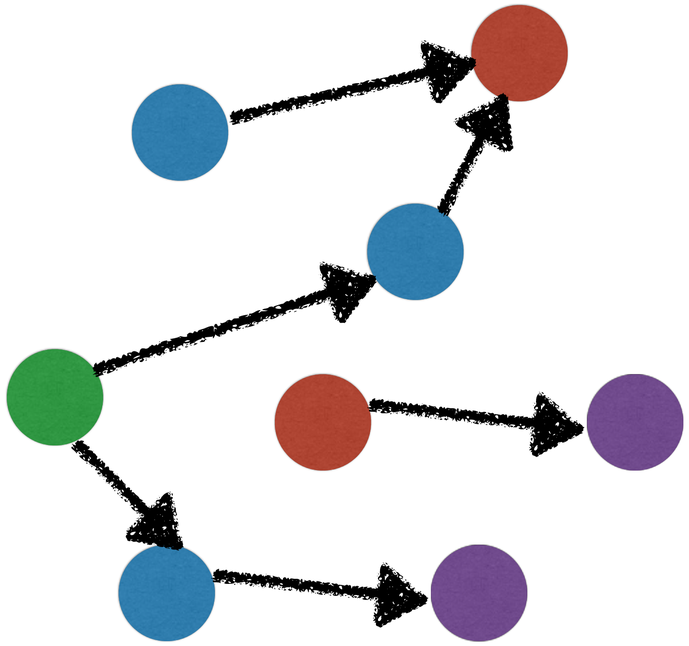
OOZIE

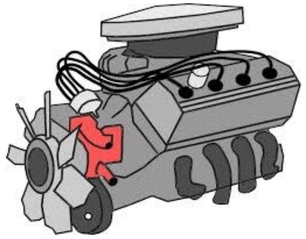


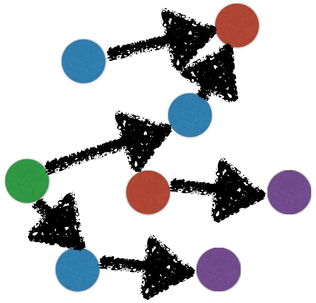


OOZIE

Each of these are individual workflows by themselves







OOZIE

You can manually process each of these workflows and execute them

That can get onerous, it's much easier to have some kind of controlling mechanism for individual workflows

OOZIE

You might want workflows to run at a certain time and frequency

First shift starts at 8am in the morning every day

OOZIE

You might want workflows to run at a certain time and frequency

Provided the input to the workflow is available

OOZIE

You might want workflows to run at a certain time and frequency

Provided the input to the workflow is available

The engine is complete, the car body is complete, mechanic is available so

fixing the engine can begin!

OOZIE

You might want workflows to run at a certain time and frequency

Provided the input to the workflow is available

This needs a Coordinator!

OOZIE

Coordinator

The Coordinator schedules the execution of a workflow at a specified time and/or a specified frequency

OOZIE

Coordinator

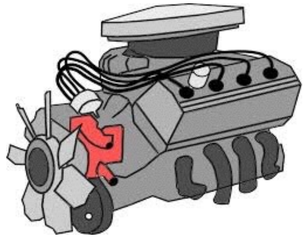
If input data is not available then the workflow is **delayed** till the data becomes available

OOZIE

Coordinator

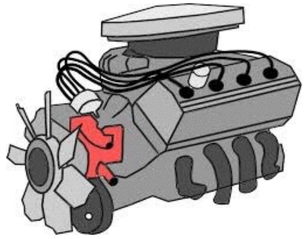
If no input data is needed then the workflow runs purely at a specified time or frequency

OOZIE

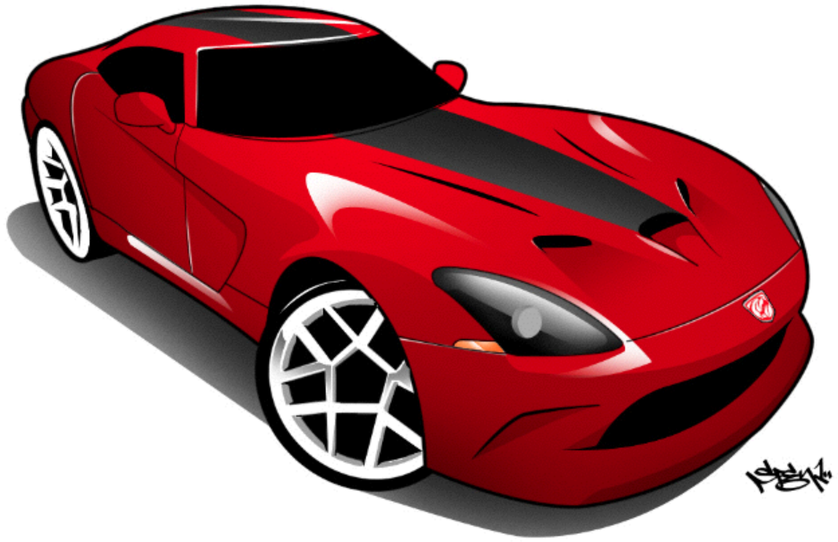


Each of these workflows can have a coordinator

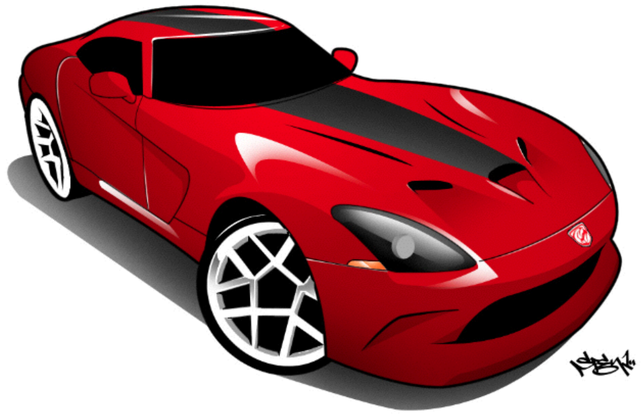
OOZIE



All these workflows come together and can build a car!



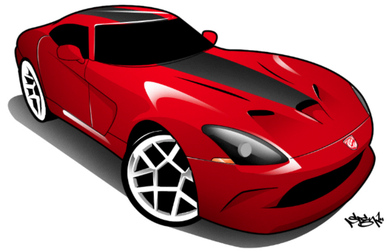
OOZIE



A car is built with a
collection of Coordinator jobs

A collection of Coordinator
jobs which can be started,
stopped and modified
together is called

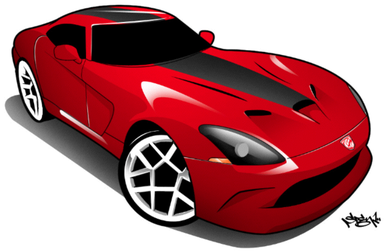
A Bundle!



OOZIE

A Bundle!

The **output** of one Coordinator job managing a workflow can be the **input** to another Coordinator job

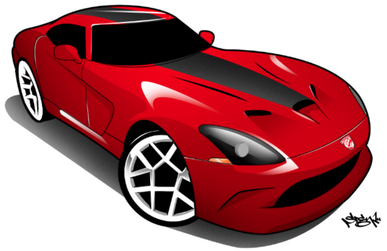


OOZIE

A Bundle!

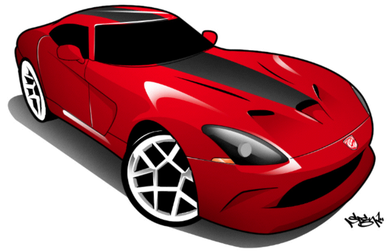
The **output** of one Coordinator job managing a workflow can be the **input** to another Coordinator job

Data Pipelines



OOZIE

Workflows, Coordinators, Bundles all come together to form the building blocks of Oozie



OOZIE

Oozie is an orchestration system
for Hadoop jobs

OOZIE

Oozie is an orchestration system for Hadoop jobs

The Hadoop eco-system is huge and gives you access to a bunch of technologies to process data

MapReduce

Hive

Pig

DistCp

Sqoop

Flume

OOZIE

Oozie is an orchestration system for Hadoop jobs

MapReduce Hive DistCp Pig Flume Sqoop

These jobs might need to be chained together to get the final output

OOZIE

Oozie is an orchestration system for Hadoop jobs

MapReduce Hive DistCp Pig Flume Sqoop

At any point you might have 1000s of Hadoop jobs running, many of which would be interdependent on one another

OOZIE

Oozie is an orchestration system for Hadoop jobs

MapReduce Hive DistCp Pig Flume Sqoop

Managing these manually or with basic scripting does not scale

OOZIE

Oozie is an orchestration system for Hadoop jobs

MapReduce Hive DistCp Pig Flume Sqoop

Oozie allows orchestration and control of such complex multi-stage Hadoop jobs

OOZIE

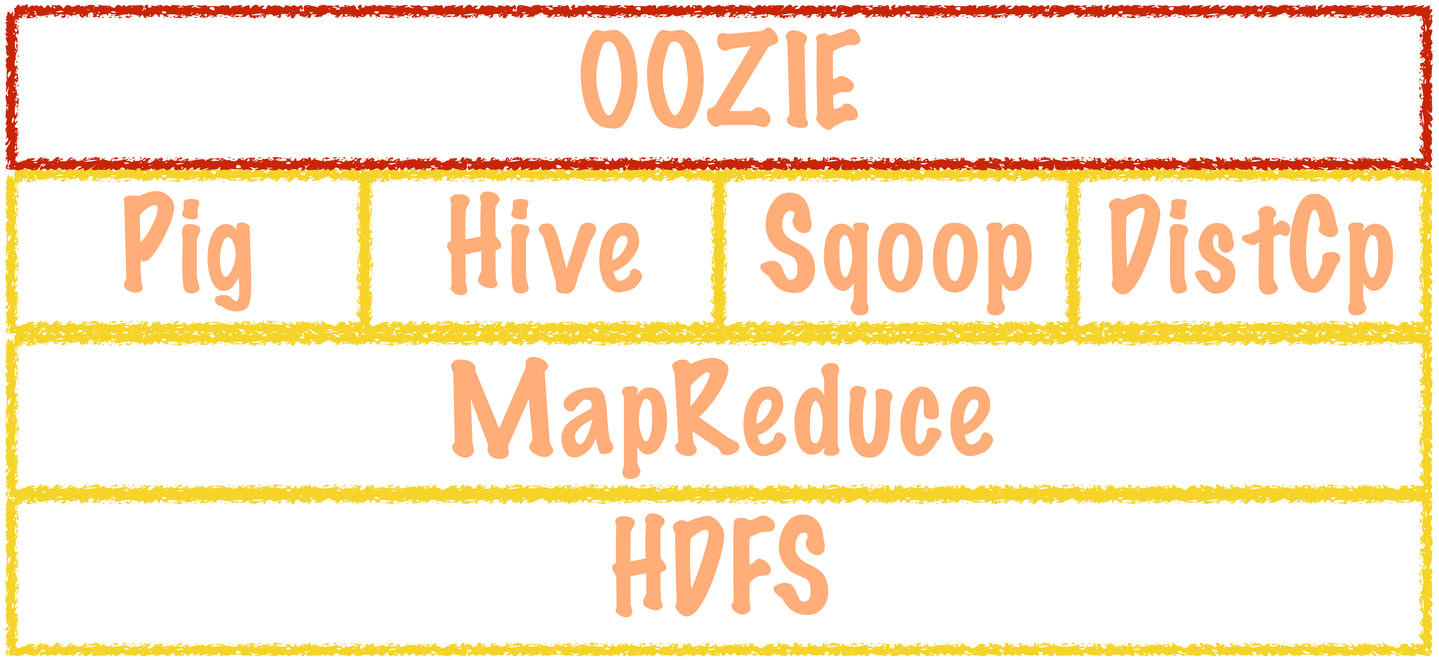
Oozie is an orchestration system for Hadoop jobs

MapReduce Hive DistCp Pig Flume Sqoop

Multi-stage Hadoop jobs can then be run as a **single Oozie job** - the Oozie job is the only thing for you to manage!

OOZIE

The Hadoop Eco-system



OOZIE

An Oozie Application

This has one file defined in XML which describes the application

It references and includes other configuration files, JARs and scripts which perform the actions

OOZIE

An Oozie Application

This has one file defined in XML
which describes the application

The application can be a workflow run
manually, a single coordinator or a
number of coordinators forming a bundle

OOZIE

An Oozie Application

This has one file defined in XML which describes the application

workflow.xml, coordinator.xml, bundle.xml files describe their corresponding applications

OOZIE

An Oozie Application

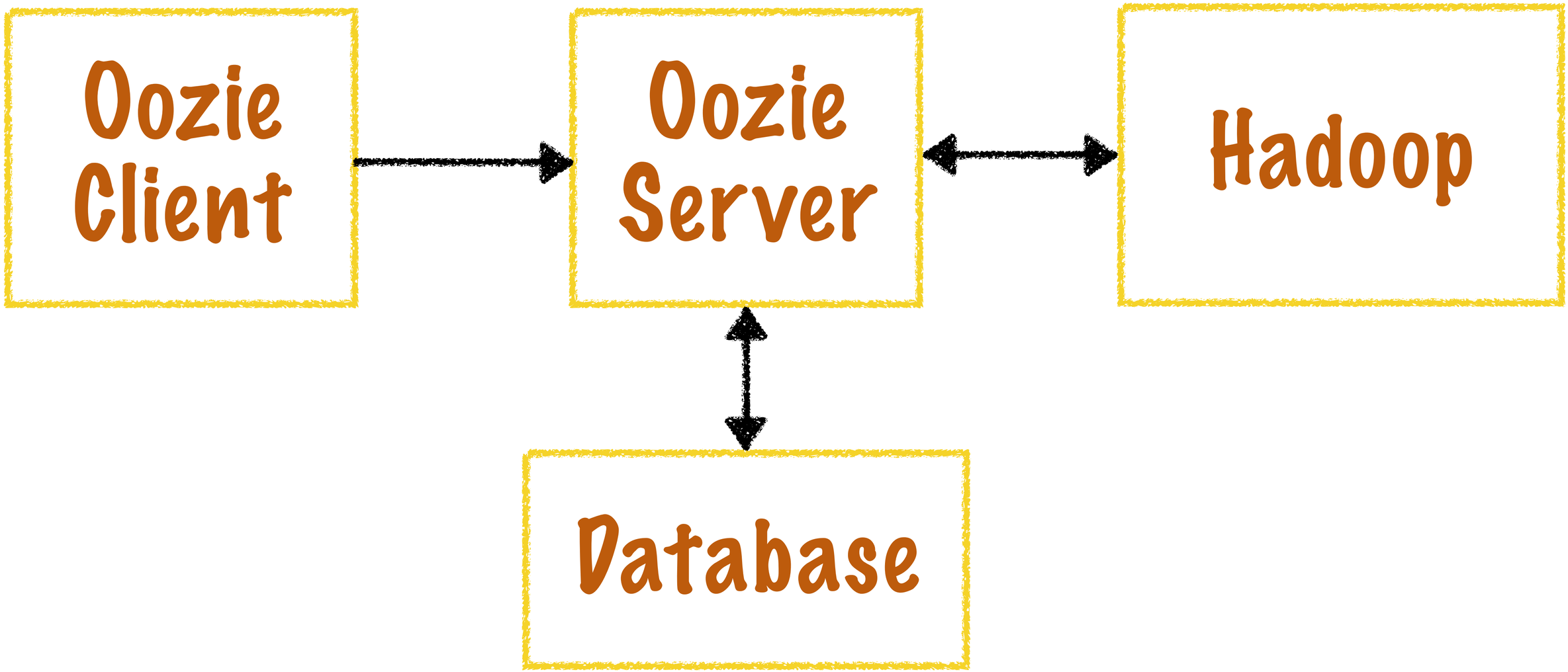
workflow.xml, coordinator.xml, bundle.xml

Oozie expects all files to be in **HDFS** before it can run

These XML files along with other files which are required for the Oozie application are **copied over to HDFS** before the job can run

OOZIE

Oozie architecture overview

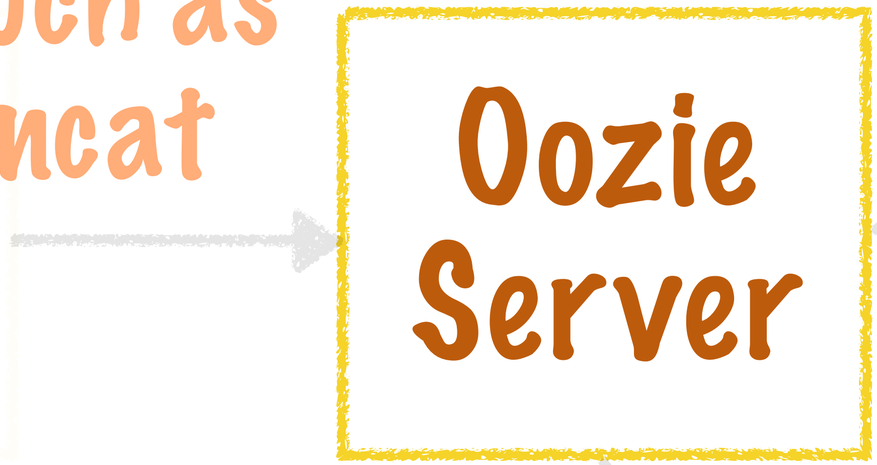


OOZIE

Oozie architecture overview

Runs in a web container such as Apache Tomcat

Manages Oozie job scheduling and execution



Is stateless, and holds job related information in the database

Oozie architecture overview

Oozie Client

Uses a command line interface to talk to the server

The server provides a REST API and a Java client so Oozie clients can be written in any language

OOZIE

Oozie architecture overview

The Oozie server is a client of Hadoop

Oozie applications read their XML from HDFS

Hadoop

They run on the Hadoop cluster

OOZIE

Oozie architecture overview

The Oozie server is stateless

All job related information is stored in the database

Oozie supports Derby, MySQL, Oracle and PostgreSQL

Database

The Oozie package comes configured with **Derby by default**

OOZIE

Oozie architecture overview