Introduction to Big Data with Apache Spark

Spark

databricks™

amplab

UC BERKELEY

Berkeley

UNIVERSITY OF CALIFORNIA

BerkeleyX

This Lecture

Programming Spark

Resilient Distributed Datasets (RDDs)

Creating an RDD

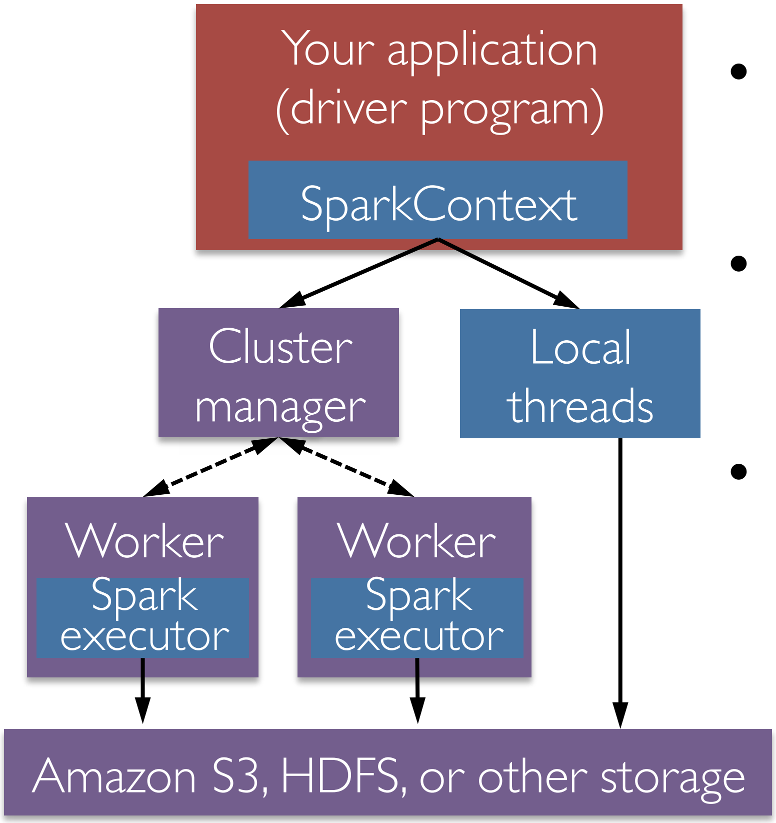
Spark Transformations and Actions

Spark Programming Model

Python Spark (pySpark)

- We are using the Python programming interface to Spark (pySpark)
- pySpark provides an easy-to-use programming abstraction and parallel runtime:
  - » "Here's an operation, run it on all of the data"
- RDDs are the key concept

Spark Driver and Workers



- A Spark program is two programs:
  - » A driver program and a workers program
- Worker programs run on cluster nodes or in local threads
- RDDs are distributed across workers

Spark Context

- A Spark program first creates a **`SparkContext`** object
  - » Tells Spark how and where to access a cluster
  - » pySpark shell and Databricks Cloud automatically create the **`sc`** variable
  - » iPython and programs must use a constructor to create a new **`SparkContext`**
- Use **`SparkContext`** to create RDDs

In the labs, we create the SparkContext for you