IPython Notebook as a Unified Data Science Interface for Hadoop

Casey Stella

Hortonworks

Spring, 2015

Table of Contents

Introduction

- I'm a Principal Architect at Hortonworks
- I work primarily doing Data Science in the Hadoop Ecosystem
- Prior to this, I've spent my time and had a lot of fun
  - Doing data mining on medical data at Explorys using the Hadoop ecosystem
  - Doing signal processing on seismic data at Ion Geophysical using MapReduce
  - Being a graduate student in the Math department at Texas A&M in algorithmic complexity theory

Data Science in Hadoop

Hadoop is a great environment for data transformation, but as a data science environment it poses challenges.

Data Science in Hadoop

Hadoop is a great environment for data transformation, but as a data science environment it poses challenges.

- A single system where both data transformation and data science algorithms can be expressed naturally can be a challenging line to toe.

Data Science in Hadoop

Hadoop is a great environment for data transformation, but as a data science environment it poses challenges.

- A single system where both data transformation and data science algorithms can be expressed naturally can be a challenging line to toe.
- The popular languages of data science with mature external libraries do not coincide with the JVM languages.

Data Science in Hadoop

Hadoop is a great environment for data transformation, but as a data science environment it poses challenges.

- A single system where both data transformation and data science algorithms can be expressed naturally can be a challenging line to toe.
- The popular languages of data science with mature external libraries do not coincide with the JVM languages.
- A system to represent the output of data science and analysis, summary analysis and visualizations, can often are either limited in scope of capabilities or require extensive custom coding.

Data Science in Hadoop

Hadoop is a great environment for data transformation, but as a data science environment it poses challenges.

- A single system where both data transformation and data science algorithms can be expressed naturally can be a challenging line to toe.
- The popular languages of data science with mature external libraries do not coincide with the JVM languages.
- A system to represent the output of data science and analysis, summary analysis and visualizations, can often are either limited in scope of capabilities or require extensive custom coding.

A unified environment for data science is elusive, but we do have a great start with the Python bindings of Spark and IPython Notebook.

Unified Data Science Environment

What are the components of a unified data science environment?

Unified Data Science Environment

What are the components of a unified data science environment?

- A single environment supporting mixed-mode local and distributed processing.

Unified Data Science Environment

What are the components of a unified data science environment?

- A single environment supporting mixed-mode local and distributed processing. **Apache Spark**

Unified Data Science Environment

What are the components of a unified data science environment?

- A single environment supporting mixed-mode local and distributed processing. **Apache Spark**
- The ability to "reach-out" to languages with heavy data science algorithm support.

Unified Data Science Environment

What are the components of a unified data science environment?

- A single environment supporting mixed-mode local and distributed processing. **Apache Spark**
- The ability to "reach-out" to languages with heavy data science algorithm support. **PySpark**

Unified Data Science Environment

What are the components of a unified data science environment?

- A single environment supporting mixed-mode local and distributed processing. **Apache Spark**
- The ability to "reach-out" to languages with heavy data science algorithm support. **PySpark**
- Strong, seamless SQL integration.

Unified Data Science Environment

What are the components of a unified data science environment?

- A single environment supporting mixed-mode local and distributed processing. **Apache Spark**
- The ability to "reach-out" to languages with heavy data science algorithm support. **PySpark**
- Strong, seamless SQL integration. **SparkSQL**

Unified Data Science Environment

What are the components of a unified data science environment?

- A single environment supporting mixed-mode local and distributed processing. **Apache Spark**
- The ability to "reach-out" to languages with heavy data science algorithm support. **PySpark**
- Strong, seamless SQL integration. **SparkSQL**
- Ability to visualize and report summary data.

Unified Data Science Environment

What are the components of a unified data science environment?

- A single environment supporting mixed-mode local and distributed processing. **Apache Spark**
- The ability to "reach-out" to languages with heavy data science algorithm support. **PySpark**
- Strong, seamless SQL integration. **SparkSQL**
- Ability to visualize and report summary data. **IPython Notebook**

Apache Spark

Apache Spark is an alternative computing system which can run on Yarn and provides

- An Elegant, Rich and Usable Core API
- An Expansive set of ecosystem libraries built around the Core API
- Hive compatibility via SparkSQL
- Mature Python support for both core APIs as well as the spark ecosystem projects

Spark: Core Ideas

Core API facilitates expressing algorithms in terms of transformations of distributed datasets

- Datasets are Distributed and Resilient (so named RDDs)
- Datasets are automatically rebuilt on failure
- Datasets have configurable persistence
- Transformations are parallel (e.g. map, reduceByKey, filter)
- Transformations support some relational primitives (e.g. join, cartesian product)

PySpark: Python Bindings

In addition to Java and Scala, Spark has solid integration with Python:

- Supports the standard Python interpreter
- There is Python support for the Spark core APIs and most ecosystem APIs, such as MLLib.
- IPython Notebook support comes out of the box

Spark: SQL Integration

The Spark component which lets you query structured data in Spark using SQL is called Spark SQL

- Has integrated APIs in Python, Scala and Java
- Allows you to integrate Spark Core APIs with SQL
- Provides Hive metastore integration so that data managed in Hive can be seamlessly processed via Spark

Open Payments Data

Sometimes, doctors and hospitals have financial relationships with health care manufacturing companies. These relationships can include money for research activities, gifts, speaking fees, meals, or travel. The Social Security Act requires CMS to collect information from applicable manufacturers and group purchasing organizations (GPOs) in order to report information about their financial relationships with physicians and hospitals.

Let's use Python and Spark via IPython Notebook to explore this dataset on Hadoop.

Questions

Thanks for your attention! Questions?

- Code & scripts for this talk available on my github presentation page.[1]
- Find me at http://caseystella.com
- Twitter handle: @casey_stella
- Email address: cstella@hortonworks.com

[1] http://github.com/cestella/presentations/