Spark for Cloudant Analytics
Hands-on-Lab (1865)

Holger Kache
Mayya Sharipova
Tony Sun

October 27, 2016

World of Watson 2016

IBM

Please note

IBM's statements regarding its plans, directions, and intent are subject to change or withdrawal without notice and at IBM's sole discretion.

Information regarding potential future products is intended to outline our general product direction and it should not be relied on in making a purchasing decision.

The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code or functionality. Information about potential future products may not be incorporated into any contract.

The development, release, and timing of any future features or functionality described for our products remains at our sole discretion.

Performance is based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput or performance that any user will experience will vary depending upon many factors, including considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve results similar to those stated here.

Agenda

- Cloudant Spark
  - What is Cloudant?
  - What is Spark?
  - What are Jupyter notebooks?
- Hands on Lab
  - Instructions
  - Python notebook
  - Scala notebook

What is Cloudant? – DBaaS for Web and Mobile Applications

Cloudant delivers a fully-managed database in service to the Analytics, App, and API economy

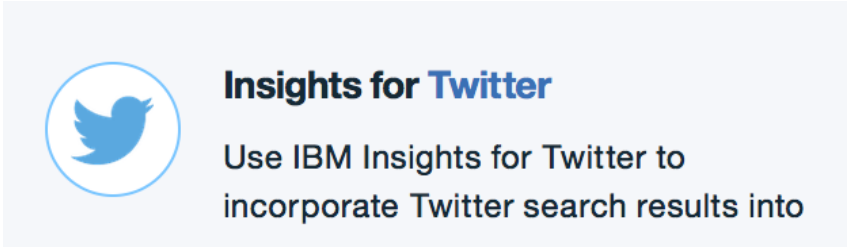
IBM **Cloudant**®

A fully-managed NoSQL database layer that can be **developed & deployed in days**

Spark Integration (Spark SQL)

- Operational NoSQL JSON store
- Master-less architecture for maximum **scalability** & **availability**
- Advanced APIs
  - REST (HTTPS) API
  - Replication & synchronization
  - Geo-load balancing
  - Incremental MapReduce indexes
  - Military-grade Geospatial indexes
  - Lucene full-text search
- Offline access to mobile apps & data
- Hybrid Cloud
  - Public | Private | Open Source | Client

JSON Documents



#Filter

- Documents are stored in the popular JSON format with a flexible schema
- A database is a logical collection of documents, with single set of access permissions
- Cluster can hold any number of databases

What is Apache spark ?

Spark is an **open** source

in-memory

computing framework for

distributed data processing and

iterative analysis

on **massive** data volumes

Key Reasons for the Interest in Spark

Performant

- In-memory architecture greatly reduces disk I/O
- Anywhere from **20-100x faster** for common tasks

Productive

- **Concise and expressive syntax**, especially compared to prior approaches
- **Single programming model** across a range of use cases and steps in data lifecycle
- **Integrated with common programming languages** – Java, Python, Scala, R
- **New tools** continually reduce skill barrier for access (e.g. SQL for analysts)

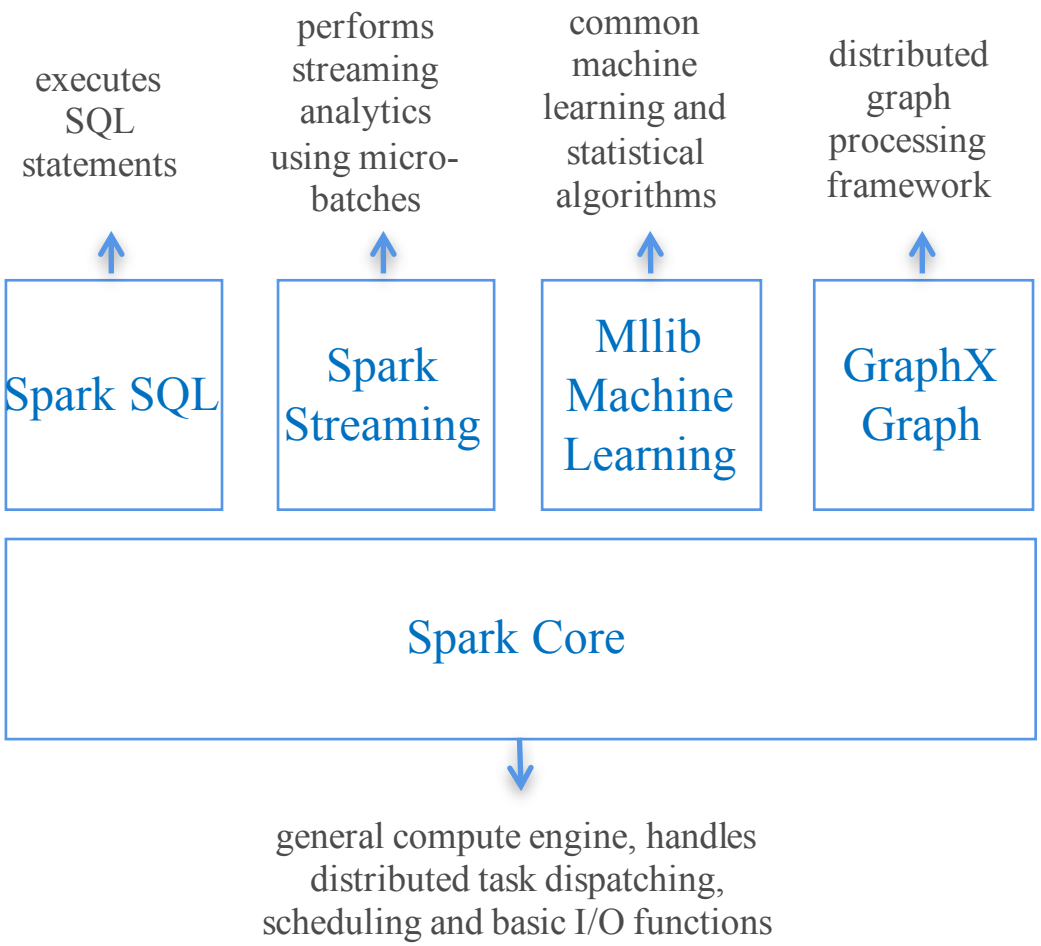
Leverages existing investments

- Works well within **existing Hadoop ecosystem**

Improves with age

- **Large and growing community** of contributors continuously improve full analytics stack and extend capabilities

Spark Core Libraries

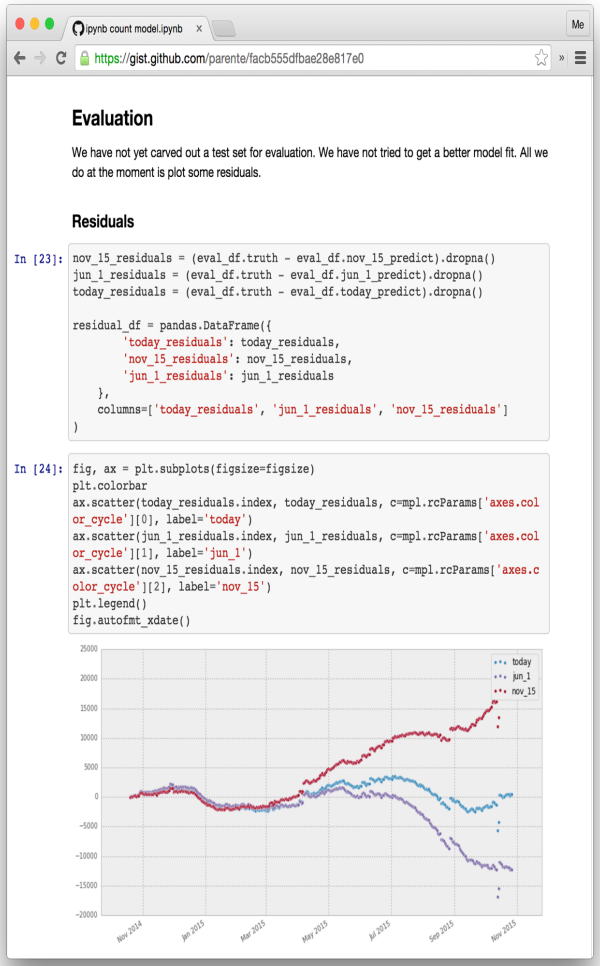


What is a Notebook?

Text, Annotations

Code, Data

Visualizations, Widgets, Output



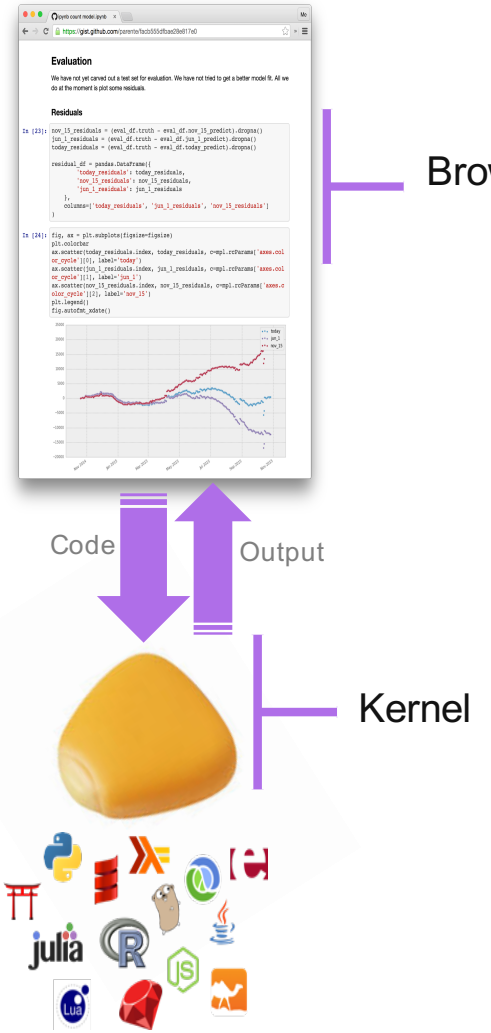
- Web based UI for running apache spark console commands
- Easy, no install spark accelerator
- Best way to start working with Spark

What is Jupyter?

with a “y”, clever ah?

"Open source, interactive data science and scientific computing"

- Formerly IPython
- Large, open, growing community and ecosystem



Cloudant Spark Connector

Built for Spark versions 1.3 – 2.0

- Tested and documented for Python and Scala
- Easy to deploy on a local Spark cluster

Implements the Spark SQL framework (http://spark.apache.org/sql/) and Spark Streaming frameworks (http://spark.apache.org/streaming/) with Cloudant specific implementations to

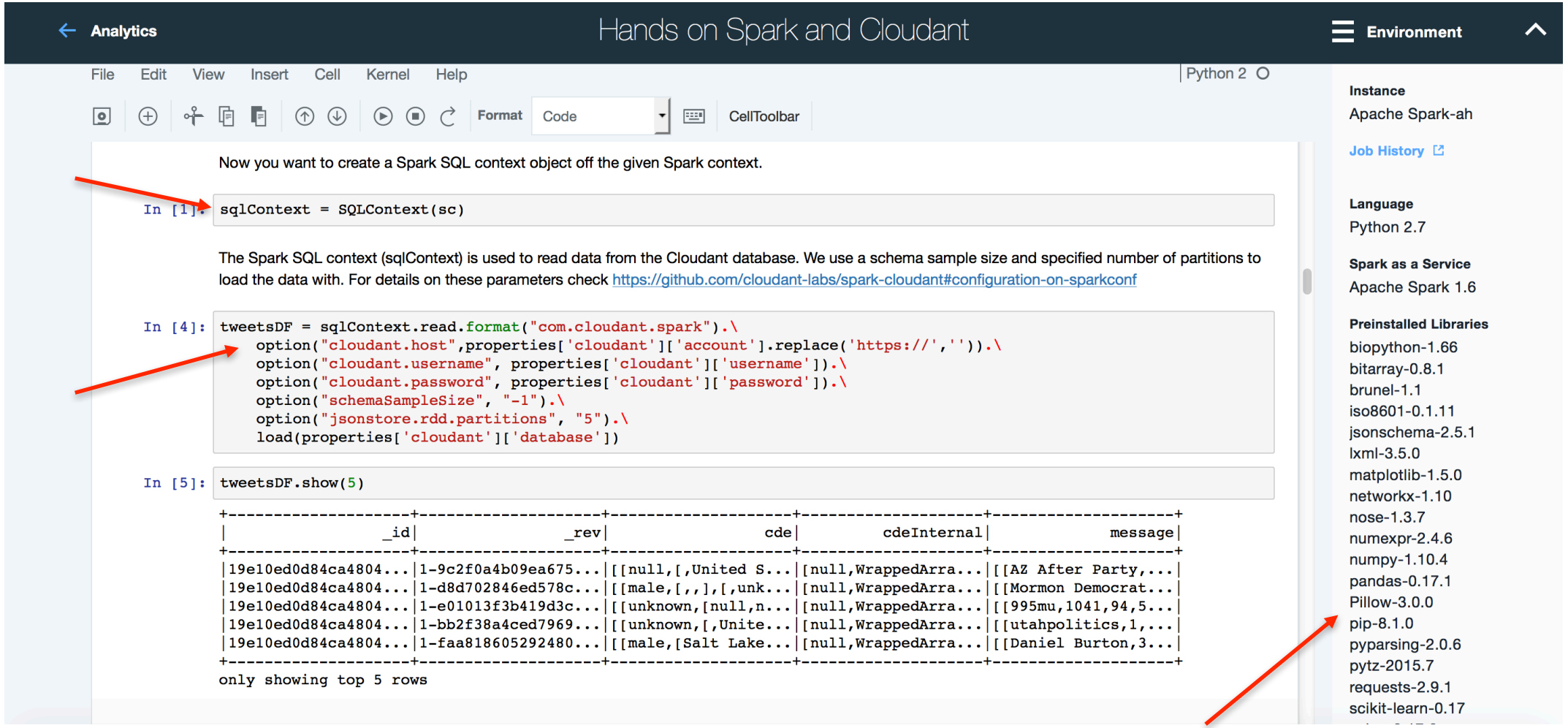
- connect to accounts
- query databases and indexes
- write to databases
- RDDs for analytics in Spark Core
- Dstreams for streaming analytics with Spark Streaming (in Scala only)

Pre-linked with the Spark-as-a-Service on Bluemix (currently at Spark version 1.6)

- available for Jupyter notebooks written in Scala and Python
- available on spark-packages.org for easy load in external Spark deployments

SparkPackages https://spark-packages.org/package/cloudant-labs/spark-cloudant

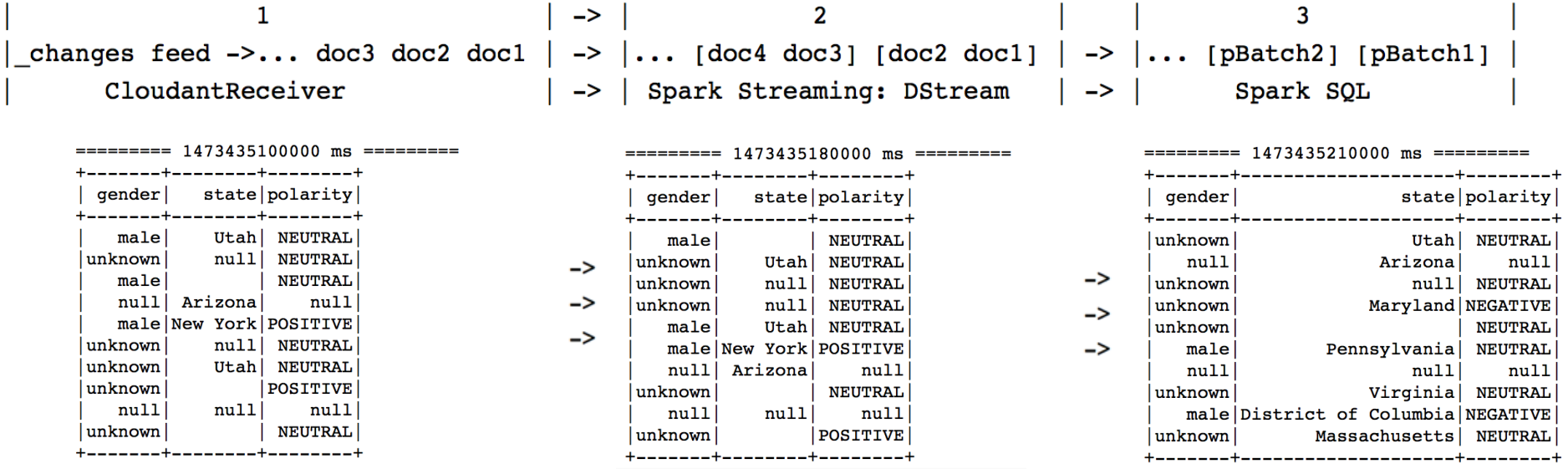
Cloudant & Spark in Jupyter notebooks



Spark Streaming through micro-batching

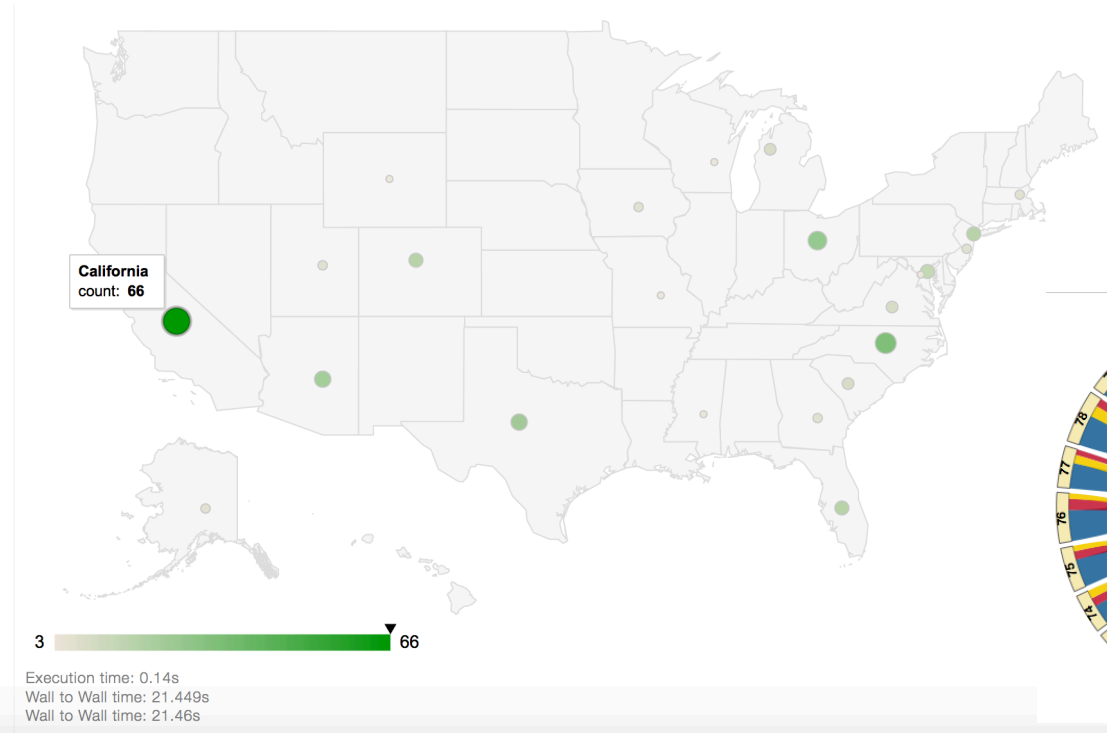
Our processing pipeline goes through three stages:

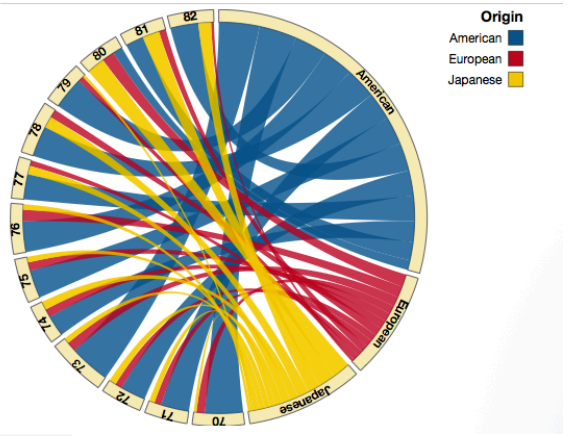
1. *_changes feed* is streamed from the given Cloudant database using `CloudantReceiver`. `CloudantReceiver` will receive _changes feed of the database, extract individual JSON documents from the feed, and store these documents in Spark's memory for processing by Spark Streaming.
2. Spark Streaming will break up this continous stream of documents into batches. Each batch is a separate RDD, and in our case represents a set of documents collected within 10 secs window. This sequence of batches, or sequence of RDDs is what is called a discretized stream or DStream.
3. Each RDD of the DStream is processed using Spark SQL.

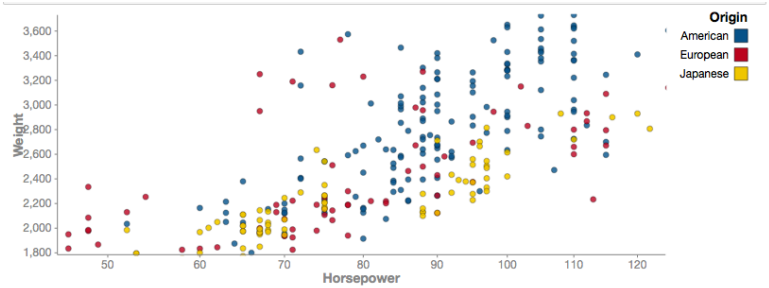


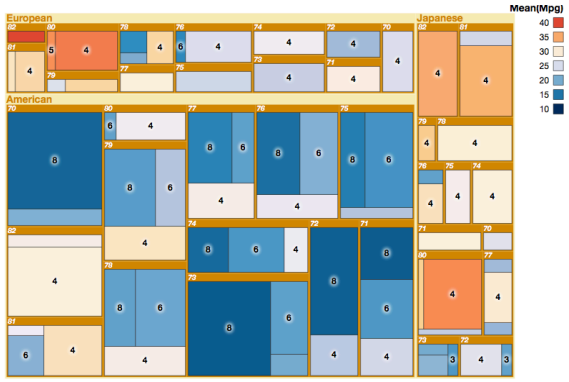
Visualization Libraries

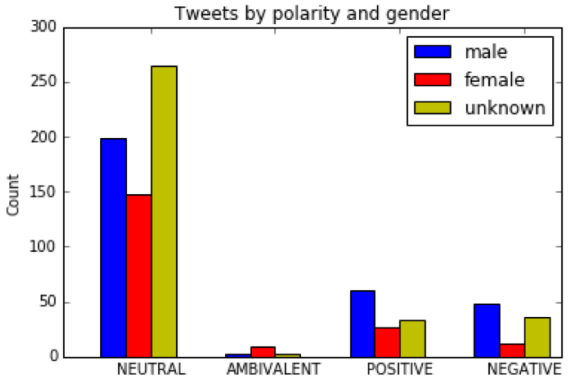
matplotlib
d3py
brunel
…











Learning Resources

Data Science Experience:

- Data Science Experience: http://datascience.ibm.com/

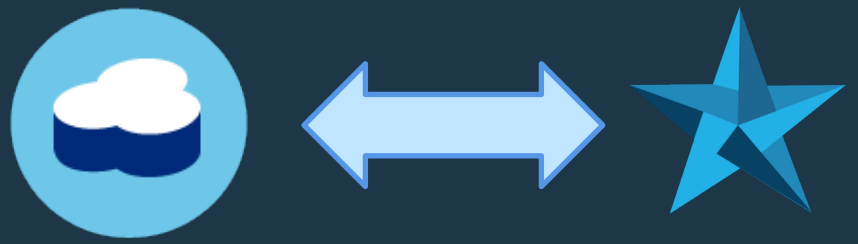
Cloudant:

- Blogs: https://cloudant.com/blog/
- Docs: https://docs.cloudant.com/
- Developer Works: http://www.ibm.com/developerworks/topics/cloudant/

Cloudant Spark

- Announcement blog: https://developer.ibm.com/clouddataservices/2016/03/09/introducing-spark-cloudant-connector/
- Scala tutorial: https://developer.ibm.com/clouddataservices/docs/cloudant/integrate/load-cloudant-data-in-apache-spark-using-scala/
- Python tutorial: https://developer.ibm.com/clouddataservices/docs/cloudant/integrate/load-cloudant-data-in-apache-spark-using-a-python-notebook/

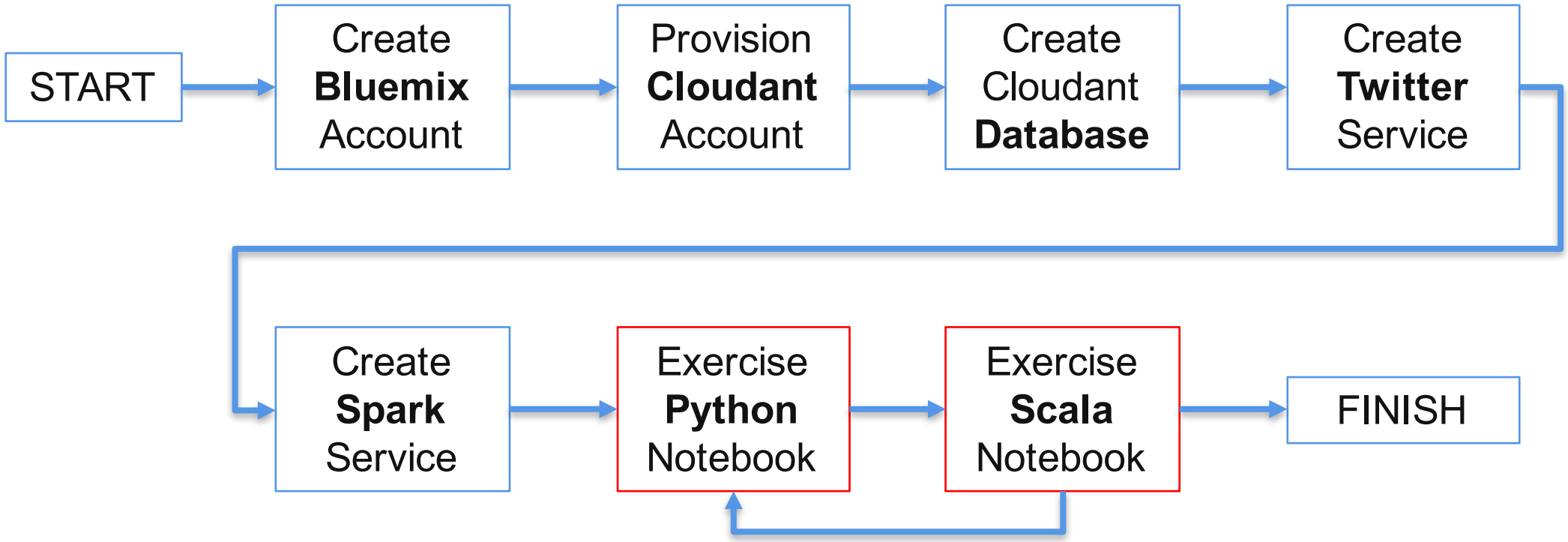
Hands on Lab



- Instructions
- Bluemix, Cloudant, Spark
- Notebooks: Python, Scala

Instructions

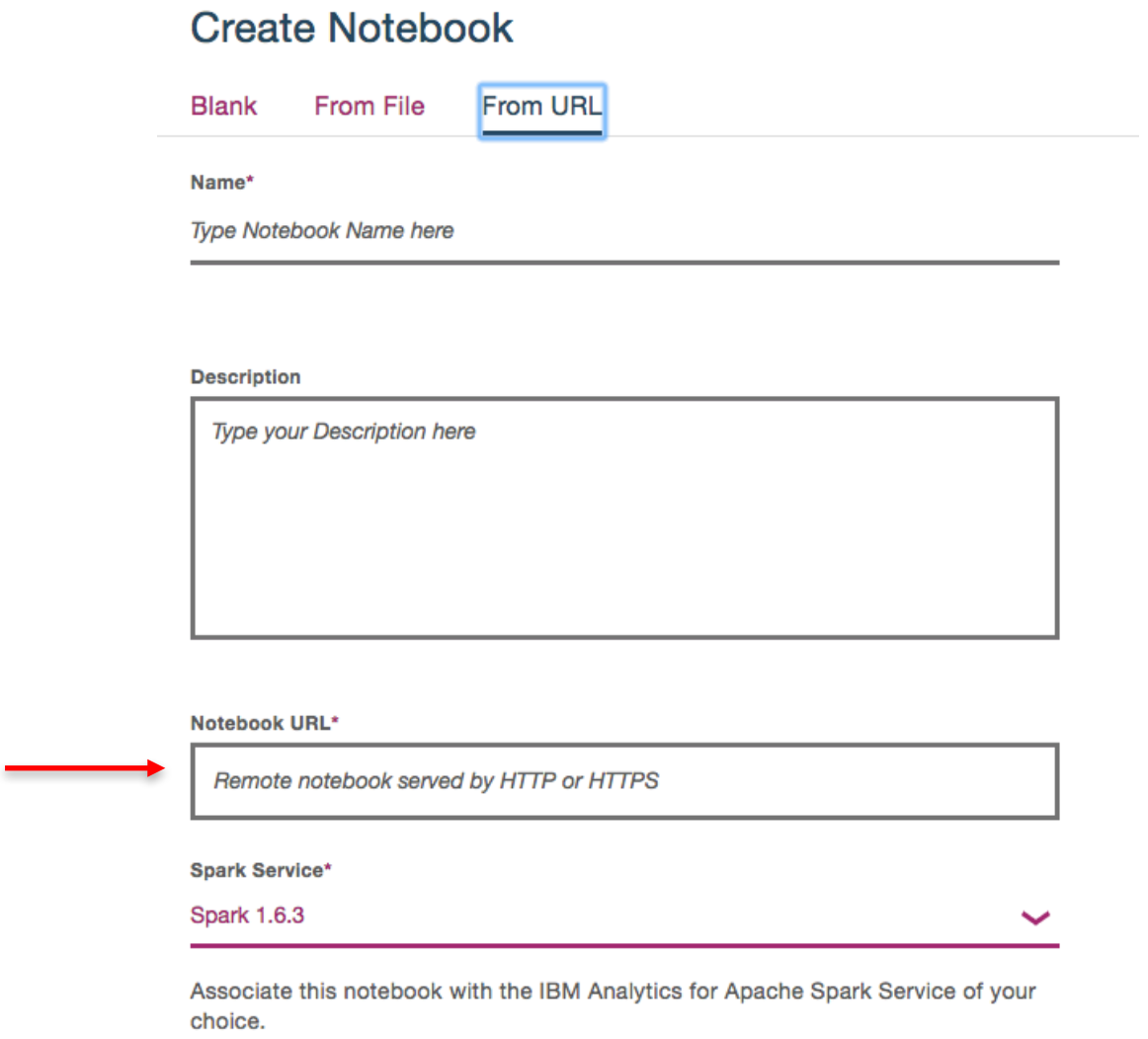
- Bring your own email ID for 30-day trial accounts
- Skip setup steps with existing accounts



Exercise Python Notebooks

This github raw file format is important!

https://raw.githubusercontent.com/cloudant-labs/spark-cloudant/master/tutorials/wowPython.ipynb



Exercise Python Notebook

```
properties = {
            'twitter': {
                'restAPI': 'https://xxx@cdeservice.mybluemix.net/api/v1/messages/search',
                'username': 'xxx',
                'password': 'xxx'
                },
            'cloudant': {
                'account':'https://xxx.cloudant.com',
                'username':'xxx',
                'password':'xxx',
                'database':'tweets'
                }
            }
```

This is for you to complete

```
query = "#election2016"
count = 300
```

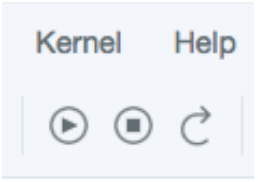
Your choice here but recommend < 3K for good performance!

```
In [6]: TtC = TwitterToCloudant()
        TtC.count = count

        TtC.query_twitter(properties, None, query, 0)
```

```
('https://5e2d04c1-cbcd-4159-901d-229e5a8d7054:JoOpVsIDMq@cdeservice.mybluemix.net/api/v1/mes
sages/search', '#election2016')
GET: Tweets from https://5e2d04c1-cbcd-4159-901d-229e5a8d7054:JoOpVsIDMq@cdeservice.mybluemi
x.net/api/v1/messages/search
Got 200 response
```

If you don't get output, restart the kernel!



Exercise Scala Notebook

Again, github raw file URL!

https://raw.githubusercontent.com/cloudant-labs/spark-cloudant/master/tutorials/wowScala.ipynb

```
val properties = Map(
    "cloudant.host"-> "ACCOUNT.cloudant.com",
    "cloudant.username"-> "USERNAME",
    "cloudant.password"-> "PASSWORD",
    "database"-> "election2016"
)
```

Configuration section

Process runs for 300 sec (that is 5 min :-))

```
ssc.start()
Thread.sleep(300000L)
ssc.stop(true)
```

Trigger a load from Twitter using the Python notebook during that time!

Results

https://github.com/cloudant-labs/spark-cloudant/master/tutorials/wowPython_RESULT.ipynb

https://github.com/cloudant-labs/spark-cloudant/master/tutorials/wowScala_RESULT.ipynb

```
In [35]: trending_pd.head(5)
```

Out[35]:

	CNT	TAG
0	1965	Election2016
1	965	TNTweeters
2	337	CIR
3	336	GOP
4	329	AINF

The hast tag cloud is visualized using the Brunel cloud graph.

```
In [36]: %brunel data('trending_pd') cloud color(cnt) size(cnt) label(tag) :: width=900, height=600
```


```
In [4]: ssc.start()
        Thread.sleep(300000L)
        ssc.stop(true)
```

```
========= 1473700520000 ms =========
+--------------------+--------------------+--------------------+--------------------+--------------------+
|                 _id|                _rev|                 cde|         cdeInternal|             message|
+--------------------+--------------------+--------------------+--------------------+--------------------+
|19e10ed0d84ca4804...|1-e01013f3b419d3c...|[[unknown,[null,n...|[null,WrappedArra...|[[995mu,1041,94,5...|
|19e10ed0d84ca4804...|1-faa818605292480...|[[male,[Salt Lake...|[null,WrappedArra...|[[Daniel Burton,3...|
|19e10ed0d84ca4804...|1-ef96819a7a98d1b...|[[male,[Monticell...|[null,WrappedArra...|[[Steven Kurlande...|
|19e10ed0d84ca4804...|1-9c2f0a4b09ea675...|[[null,[,United S...|[null,WrappedArra...|[[AZ After Party,...|
|19e10ed0d84ca4804...|1-d8d702846ed578c...|[[male,[,,],[,unk...|[null,WrappedArra...|[[Mormon Democrat...|
|19e10ed0d84ca4804...|1-bb2f38a4ced7969...|[[unknown,[,Unite...|[null,WrappedArra...|[[utahpolitics,1,...|
|19e10ed0d84ca4804...|1-0a880a2c012457c...|[[unknown,[Adelai...|[null,WrappedArra...|[[Kerry Seebohm,9...|
|19e10ed0d84ca4804...|1-85dfdf787bc125e...|[[unknown,[Köln,G...|[null,WrappedArra...|[[Awale Howle,162...|
|19e10ed0d84ca4804...|1-643e24d6fbda555...|[[male,[BELLE,Uni...|[null,WrappedArra...|[[DR.BROWN-DEAN,3...|
|19e10ed0d84ca4804...|1-020f864315fe3d6...|[[male,[Elizabeth...|[[true,null,false...|[[Mr. Huesken,164...|
+--------------------+--------------------+--------------------+--------------------+--------------------+
only showing top 10 rows

+-------+---------------+---------+
| gender|          state|polarity|
+-------+---------------+---------+
|unknown|           null| NEUTRAL|
|   male|           Utah| NEUTRAL|
|   male|       New York|POSITIVE|
|   null|        Arizona|     null|
|   male|               | NEUTRAL|
|unknown|           Utah| NEUTRAL|
|unknown|South Australia|NEGATIVE|
|unknown|               |POSITIVE|
|   male|       Missouri|NEGATIVE|
|   male|   Pennsylvania| NEUTRAL|
+-------+---------------+---------+
only showing top 10 rows

Current total count:1200
========= 1473700530000 ms =========
========= 1473700540000 ms =========
========= 1473700550000 ms =========
+--------------------+--------------------+--------------------+--------------------+--------------------+
|                 _id|                _rev|                 cde|         cdeInternal|             message|
+--------------------+--------------------+--------------------+--------------------+--------------------+
|39fb4703881bd46a4...|1-9f0eb0d3e8309ac...|[[unknown,[null,n...|[null,WrappedArra...|[[ekajoyce,179,25...|
|39fb4703881bd46a4...|1-d8d702846ed578c...|[[male,[,,],[,unk...|[null,WrappedArra...|[[Mormon Democrat...|
|39fb4703881bd46a4...|1-9c2f0a4b09ea675...|[[null,[,United S...|[null,WrappedArra...|[[AZ After Party,...|
|39fb4703881bd46a4...|1-bb2f38a4ced7969...|[[unknown,[,Unite...|[null,WrappedArra...|[[utahpolitics,1,...|
|39fb4703881bd46a4...|1-682d3282aa29a80...|[[unknown,[Accra,...|[null,WrappedArra...|[[Afoobu,218,621,...|
|39fb4703881bd46a4...|1-85dfdf787bc125e...|[[unknown,[Köln,G...|[null,WrappedArra...|[[Awale Howle,162...|
|39fb4703881bd46a4...|1-e01013f3b419d3c...|[[unknown,[null,n...|[null,WrappedArra...|[[995mu,1041,94,5...|
|39fb4703881bd46a4...|1-10c4d0c52932bee...|[[male,[,United S...|[[true,null,null,...|[[Chuck Nellis,33...|
|39fb4703881bd46a4...|1-7669e192db75074...|[[unknown,[NATION...|[null,WrappedArra...|[[Not On This Wat...|
|39fb4703881bd46a4...|1-faa818605292480...|[[male,[Salt Lake...|[null,WrappedArra...|[[Daniel Burton,3...|
+--------------------+--------------------+--------------------+--------------------+--------------------+
only showing top 10 rows
```

Notices and disclaimers

U.S. Government Users Restricted Rights - Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM.

Information in these presentations (including information relating to products that have not yet been announced by IBM) has been reviewed for accuracy as of the date of initial publication and could include unintentional technical or typographical errors. IBM shall have no responsibility to update this information. THIS DOCUMENT IS DISTRIBUTED "AS IS" WITHOUT ANY WARRANTY, EITHER EXPRESS OR IMPLIED. IN NO EVENT SHALL IBM BE LIABLE FOR ANY DAMAGE ARISING FROM THE USE OF THIS INFORMATION, INCLUDING BUT NOT LIMITED TO, LOSS OF DATA, BUSINESS INTERRUPTION, LOSS OF PROFIT OR LOSS OF OPPORTUNITY. IBM products and services are warranted according to the terms and conditions of the agreements under which they are provided.

IBM products are manufactured from new parts or new and used parts. In some cases, a product may not be new and may have been previously installed. Regardless, our warranty terms apply."

Any statements regarding IBM's future direction, intent or product plans are subject to change or withdrawal without notice.

Performance data contained herein was generally obtained in a controlled, isolated environments. Customer examples are presented as illustrations of how those customers have used IBM products and the results they may have achieved. Actual performance, cost, savings or other results in other operating environments may vary.

References in this document to IBM products, programs, or services does not imply that IBM intends to make such products, programs or services available in all countries in which IBM operates or does business.

Workshops, sessions and associated materials may have been prepared by independent session speakers, and do not necessarily reflect the views of IBM. All materials and discussions are provided for informational purposes only, and are neither intended to, nor shall constitute legal or other guidance or advice to any individual participant or their specific situation.

It is the customer's responsibility to insure its own compliance with legal requirements and to obtain advice of competent legal counsel as to the identification and interpretation of any relevant laws and regulatory requirements that may affect the customer's business and any actions the customer may need to take to comply with such laws. IBM does not provide legal advice or represent or warrant that its services or products will ensure that the customer is in compliance with any law.

Notices and disclaimers continued

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products in connection with this publication and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products. IBM does not warrant the quality of any third-party products, or the ability of any such third-party products to interoperate with IBM's products. IBM EXPRESSLY DISCLAIMS ALL WARRANTIES, EXPRESSED OR IMPLIED, INCLUDING BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE.

The provision of the information contained herein is not intended to, and does not, grant any right or license under any IBM patents, copyrights, trademarks or other intellectual property right.


Thank You
World of
Watson
2016
IBM