Microsoft

Cortana Analytics Workshop

Sept 10 – 11, 2015 • MSCC

New Services in the Cortana Analytics Suite

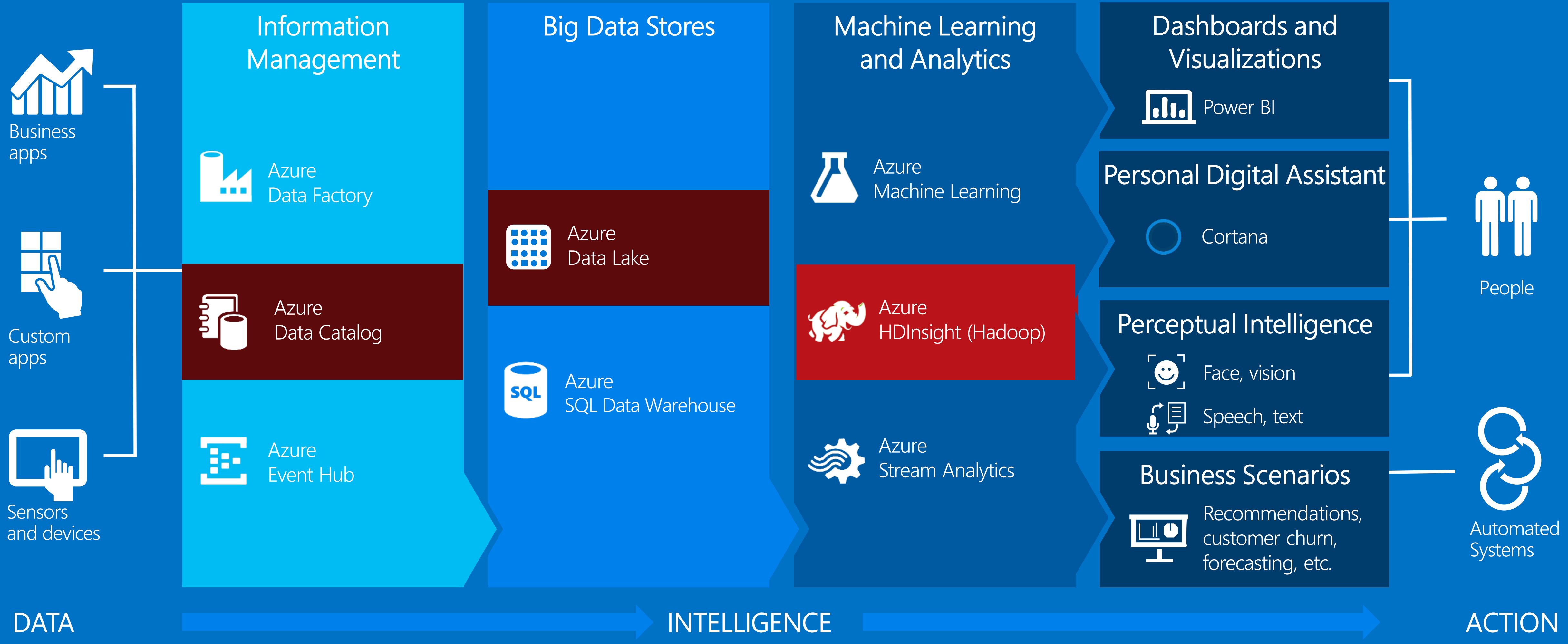
Apache Spark in HDInsight

Asad Khan
Principal Program Manager
Big Data Group

Microsoft

Cortana Analytics Suite

Big Data and Advanced Analytics



Agenda

Introduce Apache Spark

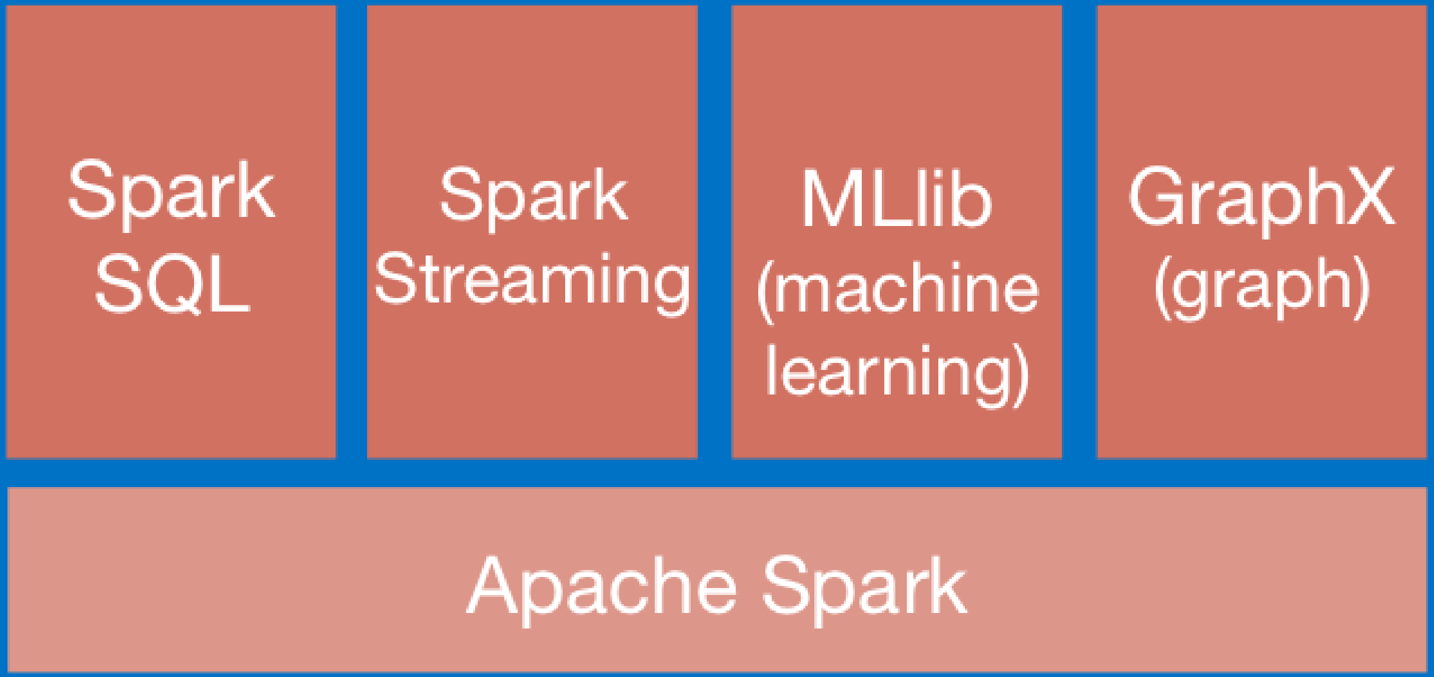
Spark in Action

Apache Spark – An Unified Framework

An unified, open source, parallel, data processing framework for Big Data Analytics

Spark Unifies:

- ☆ Batch Processing
- ☆ Real-time processing
- ☆ Stream Analytics
- ☆ Machine Learning
- ☆ Interactive SQL



https://spark.apache.org

Spark - Benefits

Performance

Using in-memory computing, Spark is considerably faster than Hadoop (100x in some tests).
Can be used for batch and real-time data processing.

Developer Productivity

Easy-to-use APIs for processing large datasets.
Includes 100+ operators for transforming.

Unified Engine

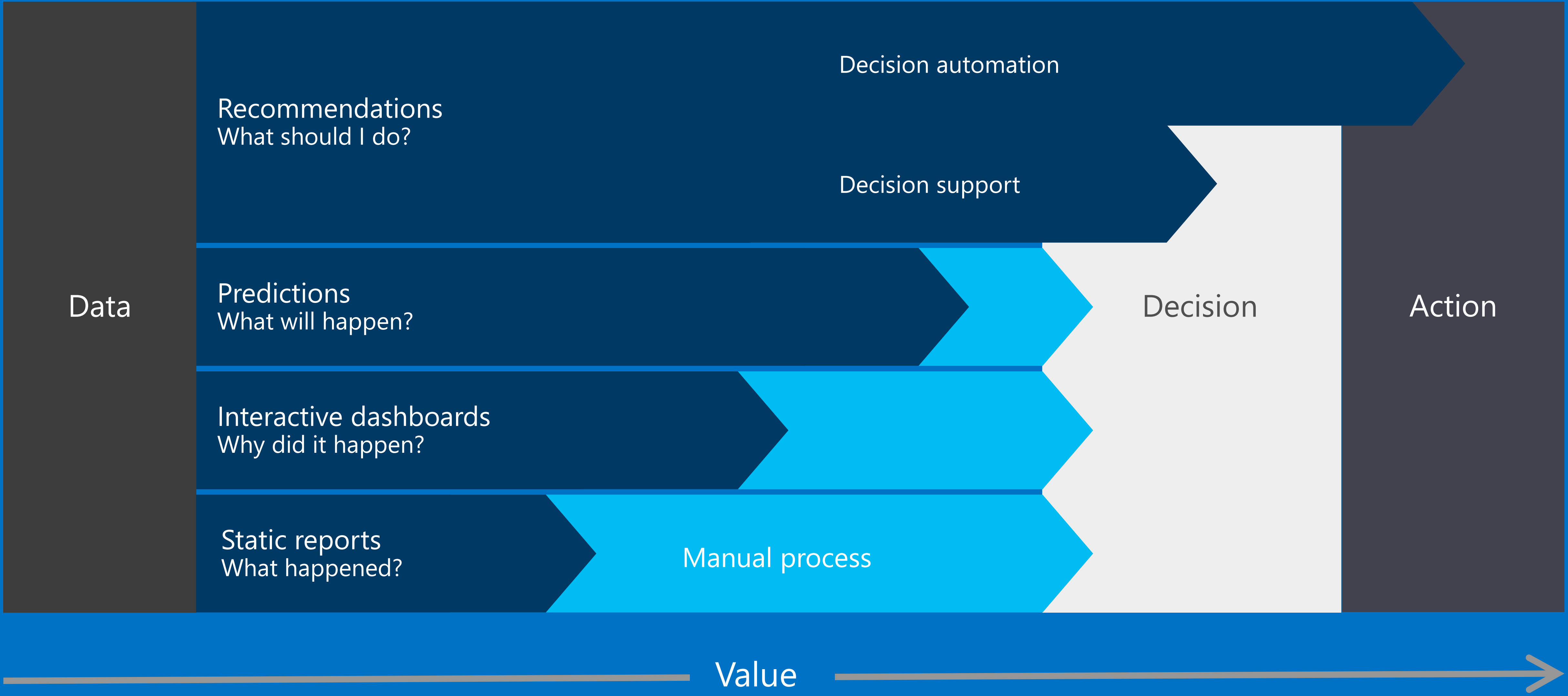
Integrated framework includes higher-level libraries for interactive SQL queries, processing streaming data, machine learning and graph processing.
A single application can combine all types of processing

Ecosystem

Spark has built-in support for many data sources such as HDFS, RDBMS, S3, Apache Hive, Cassandra and MongoDB.

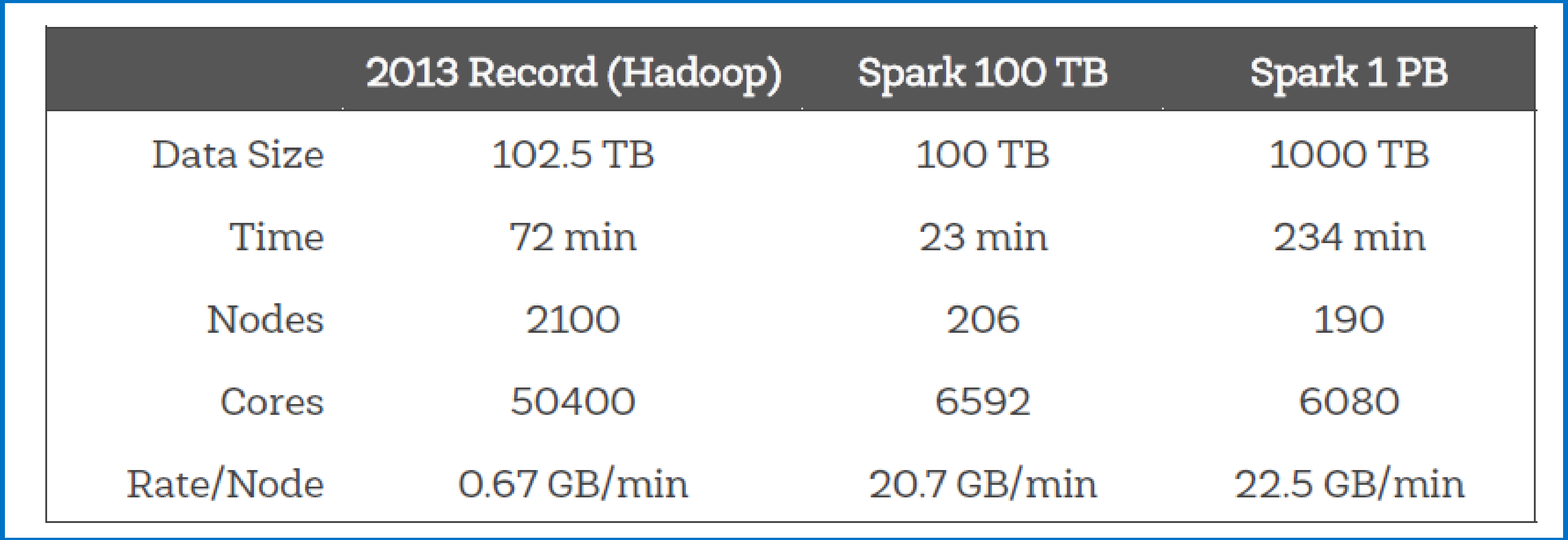
Runs on top the Apache YARN resource manager.

From data to decisions and actions



Spark is fast

Spark is the current (2014) Sort Benchmark winner.
3x faster than 2013 winner (Hadoop).

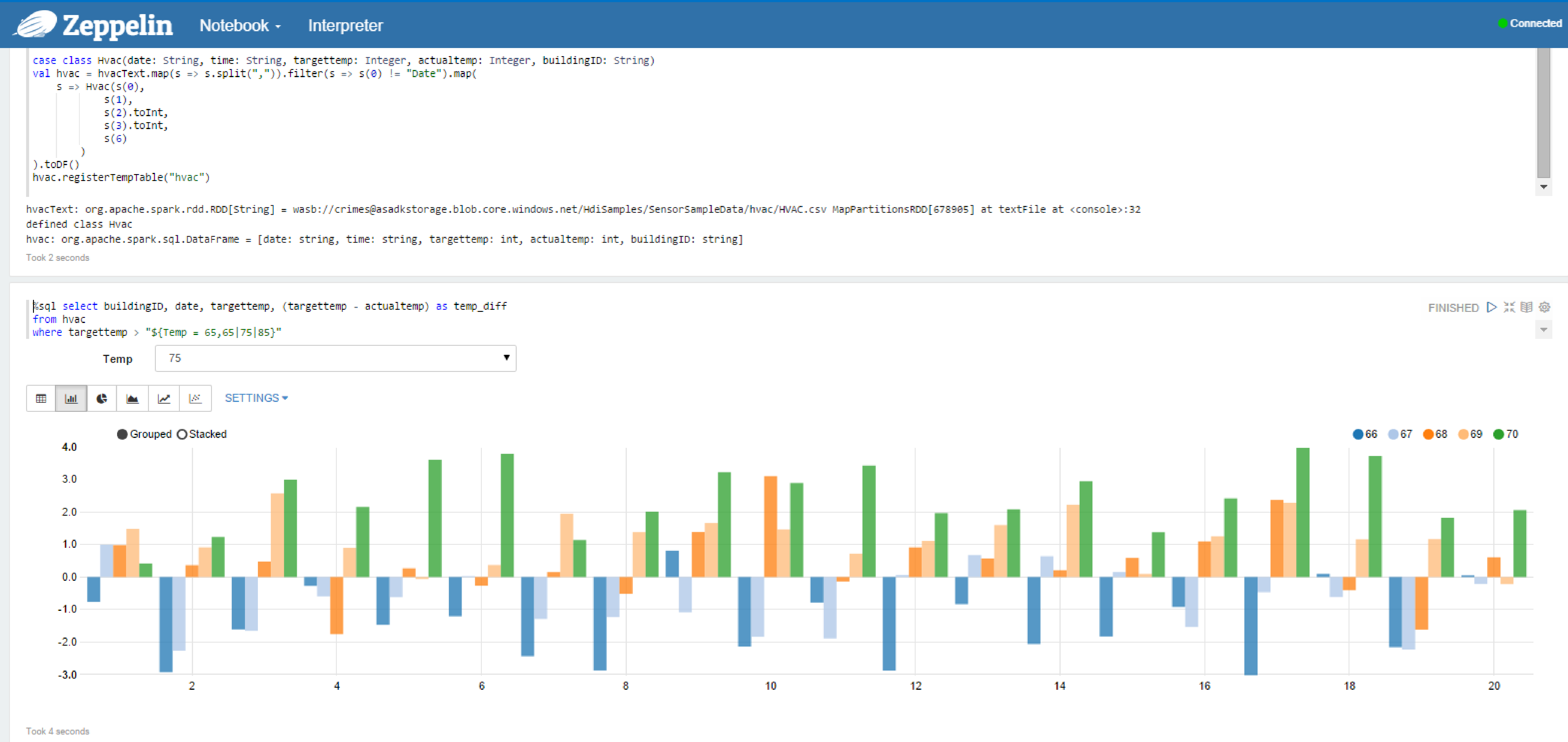


	2013 Record (Hadoop)	Spark 100 TB	Spark 1 PB
Data Size	102.5 TB	100 TB	1000 TB
Time	72 min	23 min	234 min
Nodes	2100	206	190
Cores	50400	6592	6080
Rate/Node	0.67 GB/min	20.7 GB/min	22.5 GB/min

Spark is fast not just for In-Memory but On-Disk computation as well

tinyurl.com/spark-sort

Interactive Data analysis through Zeppelin



```
case class Hvac(date: String, time: String, targettemp: Integer, actualtemp: Integer, buildingID: String)
val hvac = hvacText.map(s => s.split(",")).filter(s => s(0) != "Date").map(
    s => Hvac(s(0),
            s(1),
            s(2).toInt,
            s(3).toInt,
            s(6)
        )
).toDF()
hvac.registerTempTable("hvac")
```

```
hvacText: org.apache.spark.rdd.RDD[String] = wasb://crimes@asadkstorage.blob.core.windows.net/HdiSamples/SensorSampleData/hvac/HVAC.csv MapPartitionsRDD[678905] at textFile at <console>:32
defined class Hvac
hvac: org.apache.spark.sql.DataFrame = [date: string, time: string, targettemp: int, actualtemp: int, buildingID: string]
```

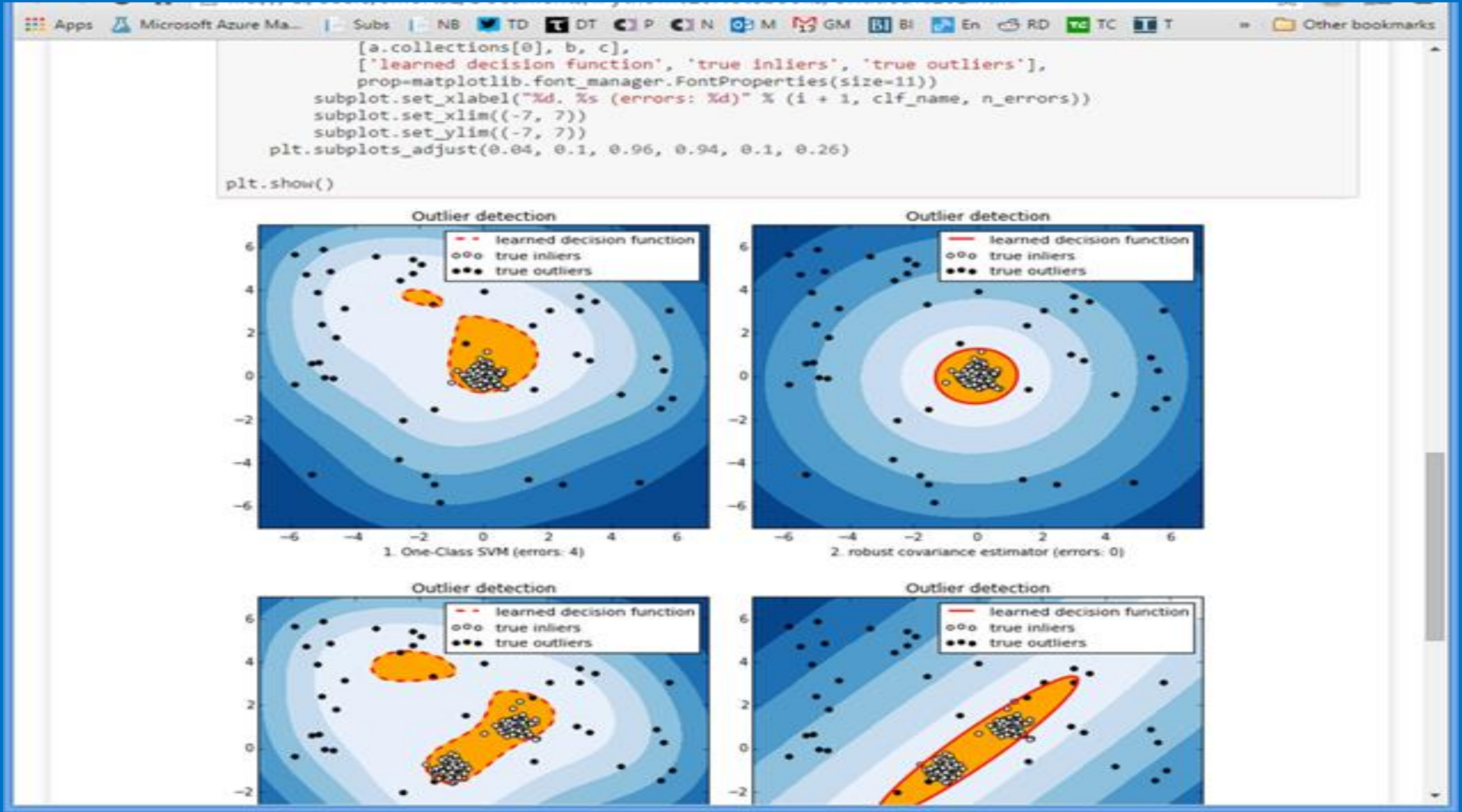
Took 2 seconds

```
%sql select buildingID, date, targettemp, (targettemp - actualtemp) as temp_diff
from hvac
where targettemp > "${Temp = 65,65|75|85}"
```

Temp: 75

Took 4 seconds

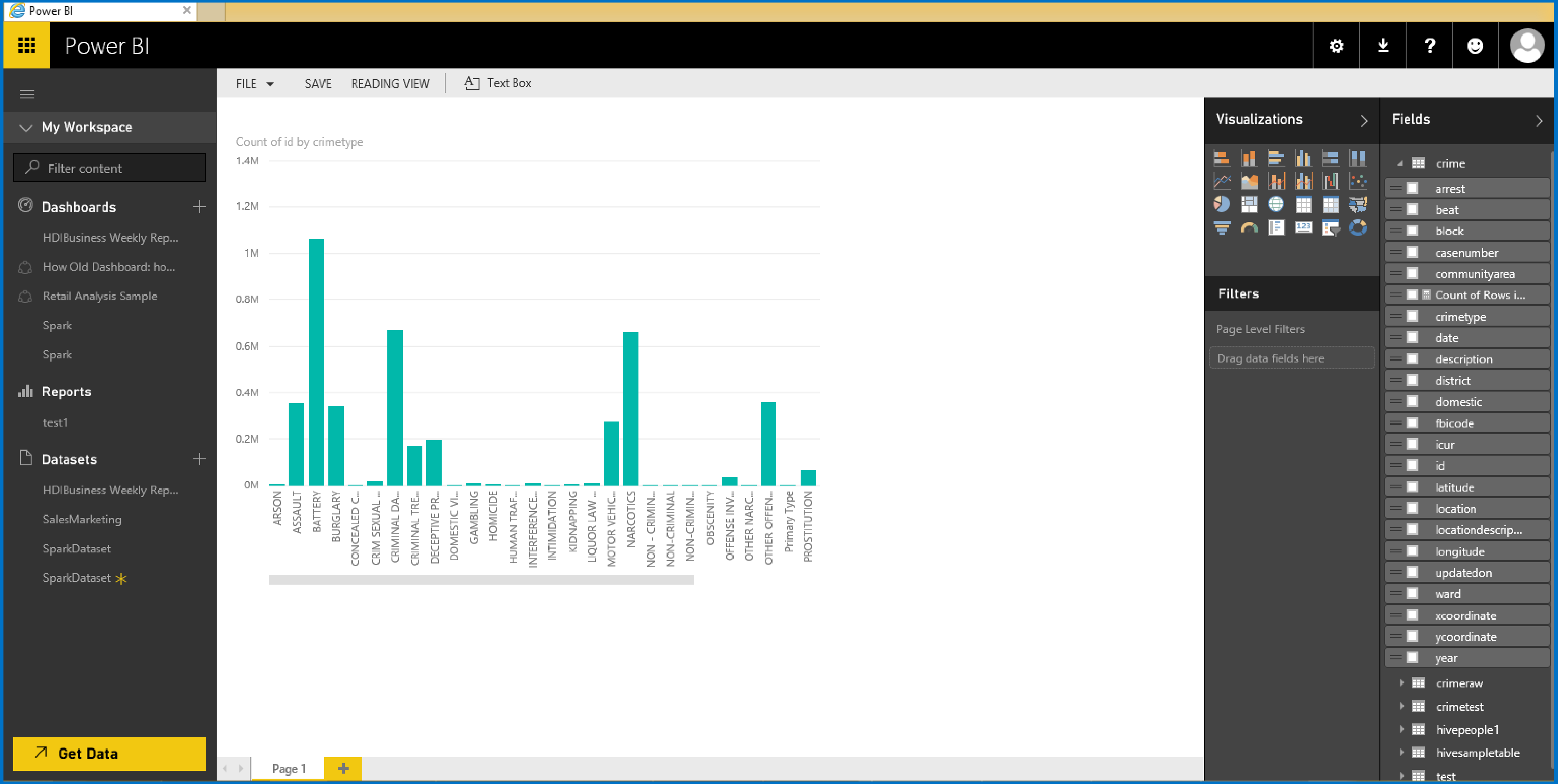
Machine learning though Juypter



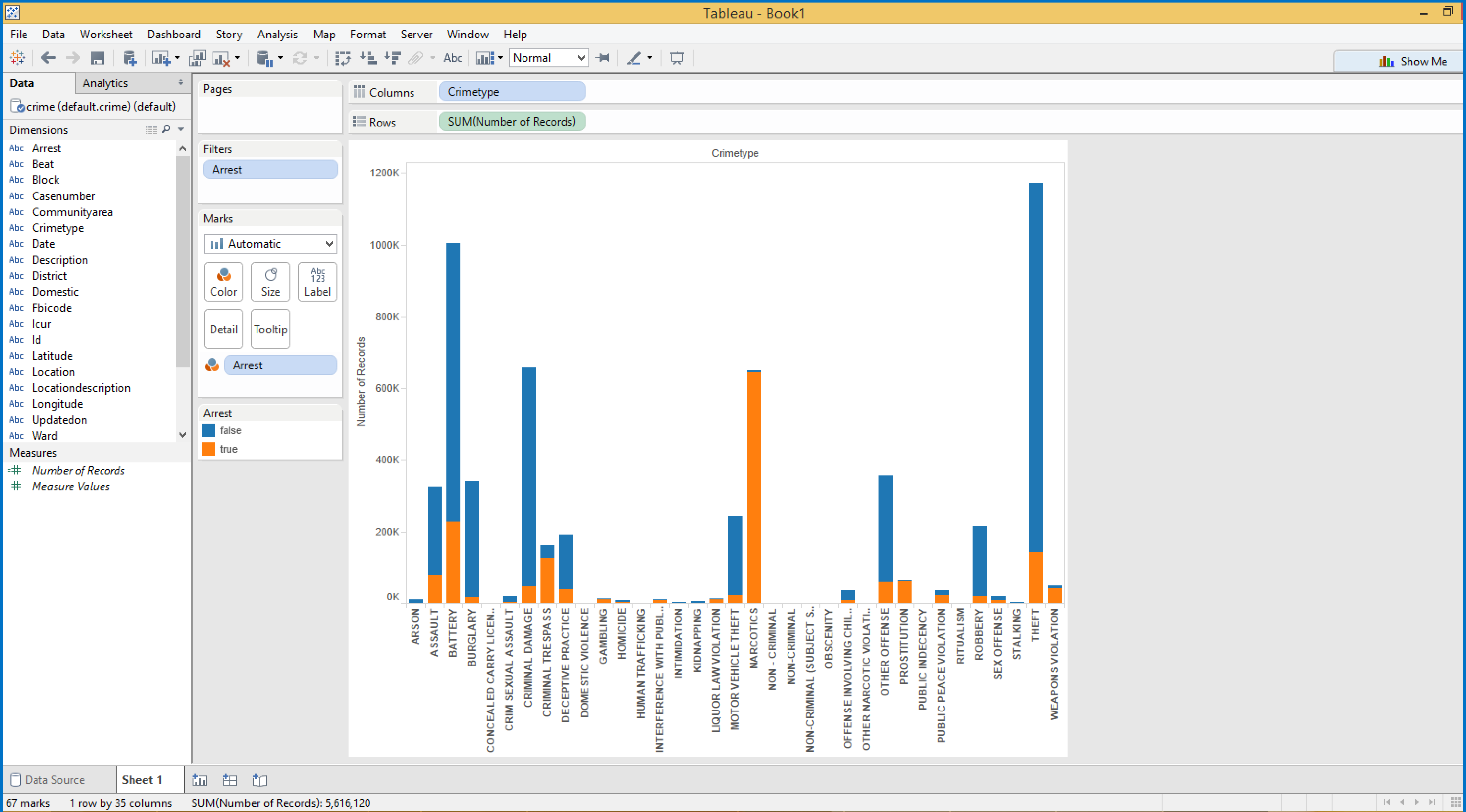
```
            [a.collections[0], b, c],
            ['learned decision function', 'true inliers', 'true outliers'],
            prop=matplotlib.font_manager.FontProperties(size=11))
        subplot.set_xlabel("%d. %s (errors: %d)" % (i + 1, clf_name, n_errors))
        subplot.set_xlim((-7, 7))
        subplot.set_ylim((-7, 7))
    plt.subplots_adjust(0.04, 0.1, 0.96, 0.94, 0.1, 0.26)

plt.show()
```

Visual exploration through BI



Visual exploration through BI



Arrest

🟦	false
🟧	true

67 marks   1 row by 35 columns   SUM(Number of Records): 5,616,120

Agenda

Introduce Apache Spark

Spark in Action

Microsoft

Contact:

asadk@microsoft.com