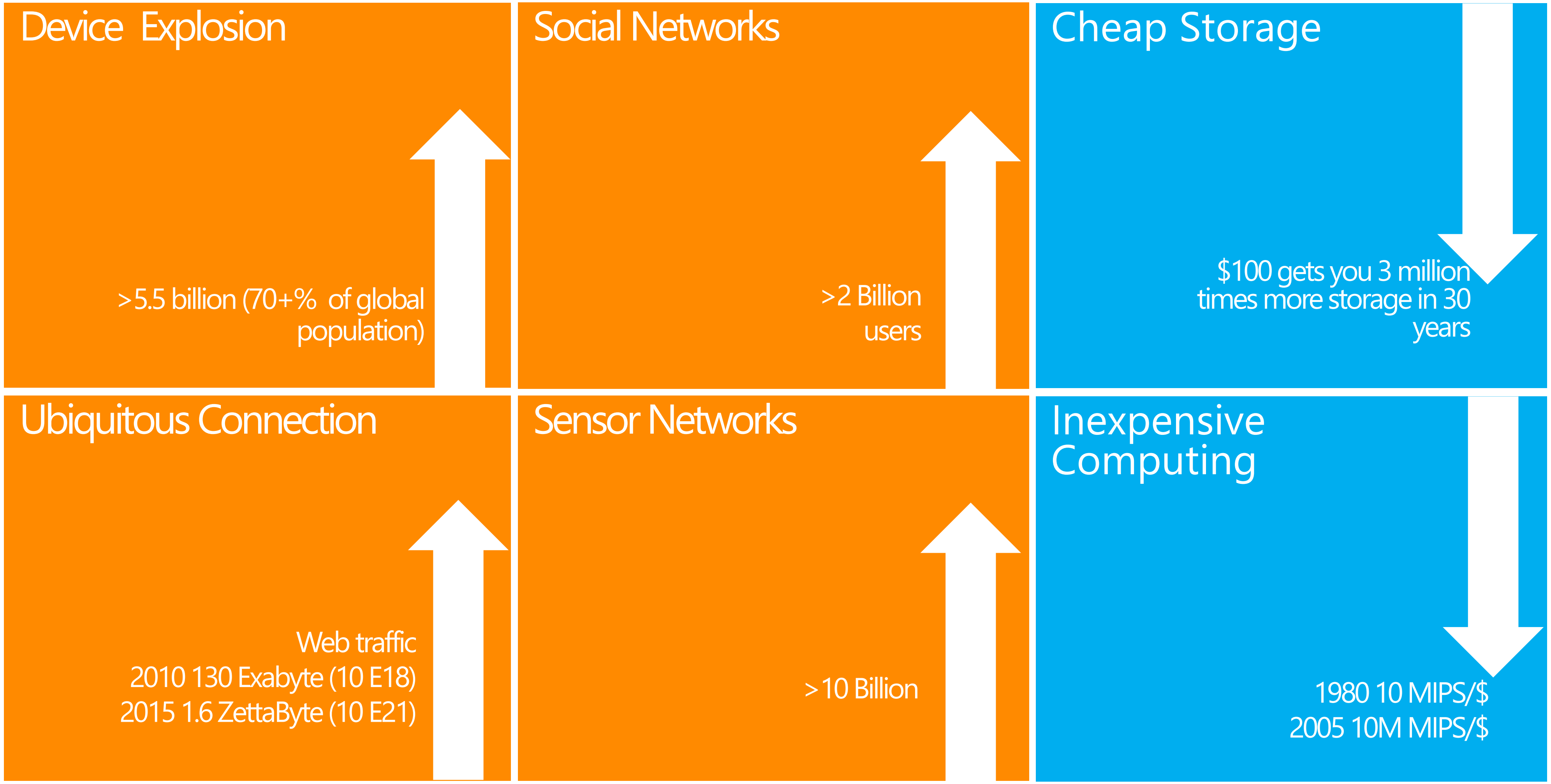
Microsoft Azure

Big-Data Analytics with Azure HDInsight

Microsoft Research

Big Data



Device Explosion

>5.5 billion (70+% of global population)

Social Networks

>2 Billion users

Cheap Storage

$100 gets you 3 million times more storage in 30 years

Ubiquitous Connection

Web traffic
2010 130 Exabyte (10 E18)
2015 1.6 ZettaByte (10 E21)

Sensor Networks

>10 Billion

Inexpensive Computing

1980 10 MIPS/$
2005 10M MIPS/$

Azure HDInsight

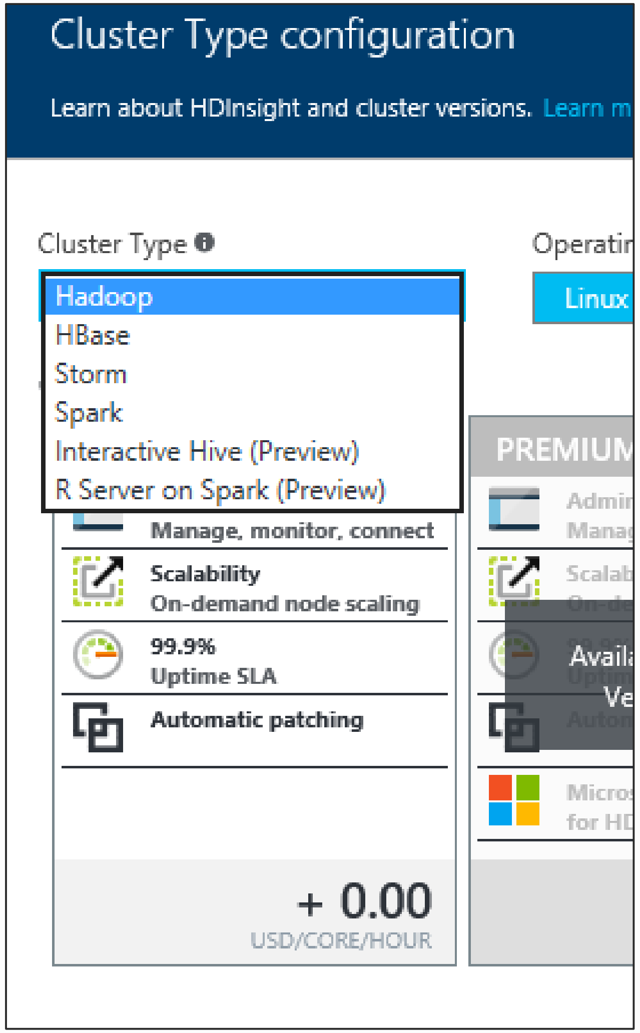
- Microsoft Azure's big-data solution using Hadoop
  - Open-source framework for storing and analyzing massive amounts of data on clusters built from commodity hardware
  - Uses Hadoop Distributed File System (HDFS) for storage
- Employs the open-source Hortonworks Data Platform implementation of Hadoop
  - Includes Hive, Pig, Storm, Spark, and more
- Integrates with popular BI tools
  - Includes Power BI, Excel, SSAS, SSRS, Tableau

Why Hadoop on Azure?

- Automatic cluster provisioning & configuration
  - Bypass an otherwise manual-intensive process
- Cluster scaling
  - Change number of nodes without deleting/re-creating the cluster
- High availability/reliability
  - Managed solution - 99.9% SLA
  - HDInsight includes a secondary head node
- Reliable and economical storage
  - HDFS mapped over Azure Blob Storage
  - Accessed through “wasb://” protocol prefix

HDInsight Cluster Types

- Hadoop: Query workloads
  - Reliable data storage, simple MapReduce
- HBase: NoSQL workloads
  - Distributed database offering random access to large amounts of data
- Apache Storm: Stream workloads
  - Real-time analysis of moving data streams
- Apache Spark: High-performance workloads
  - In-memory parallel processing



Cluster Type configuration

Learn about HDInsight and cluster versions. Learn m

Cluster Type

Hadoop

Operatir

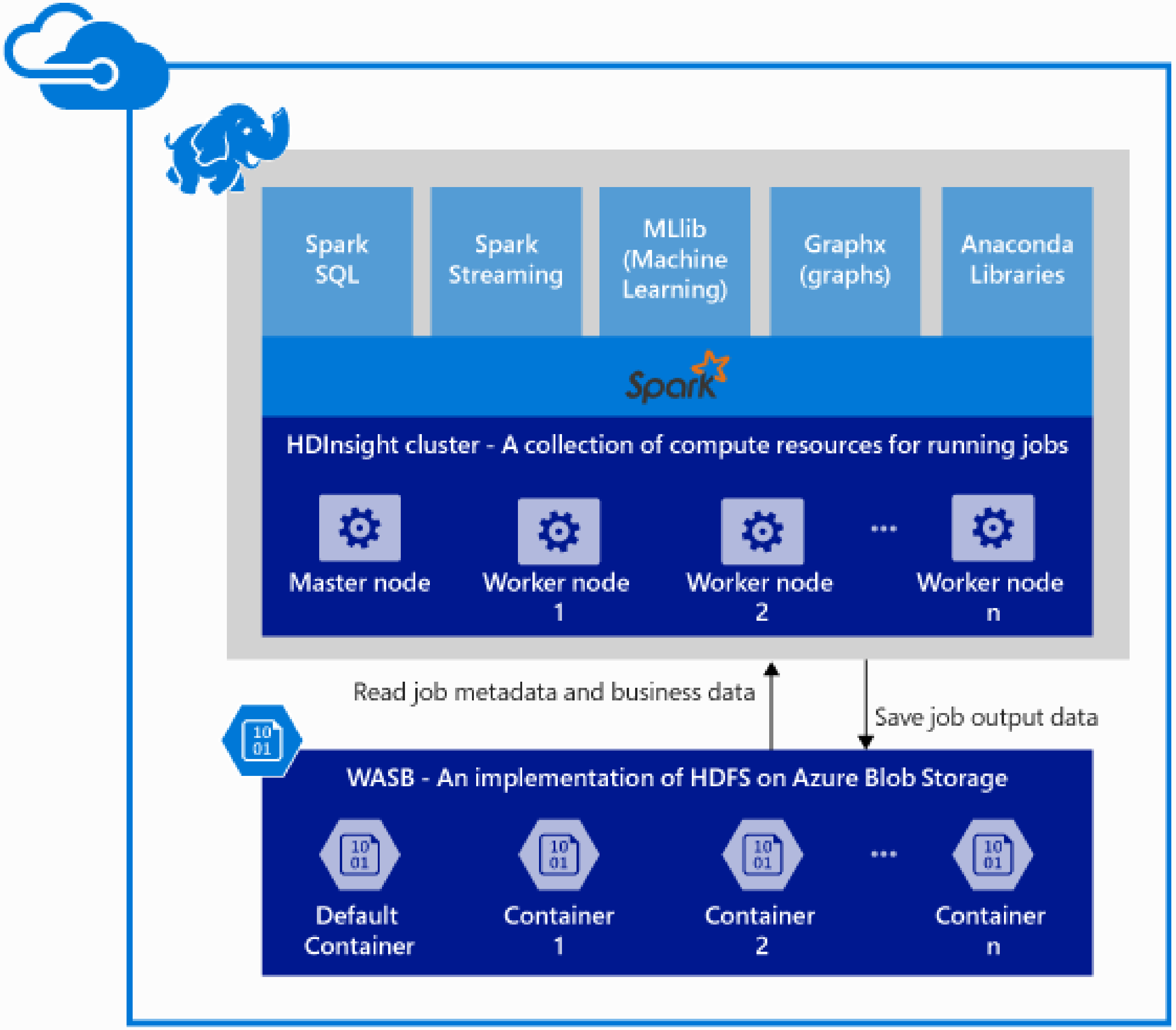
Linux

Apache Spark

- Interactive manipulation and visualization of data
  - Scala, Python, and R Interactive Shells
  - Jupyter Notebook with PySpark (Python) and Spark (Scala) kernels provide in-browser interaction
- Unified platform for processing multiple workloads
  - Real-time processing, Machine Learning, Stream Analytics, Interactive Querying, Graphing
- Leverages in-memory processing for really big data
  - Resilient distributed datasets (RDDs)
  - APIs for processing large datasets
  - Up to 100x faster than Hadoop

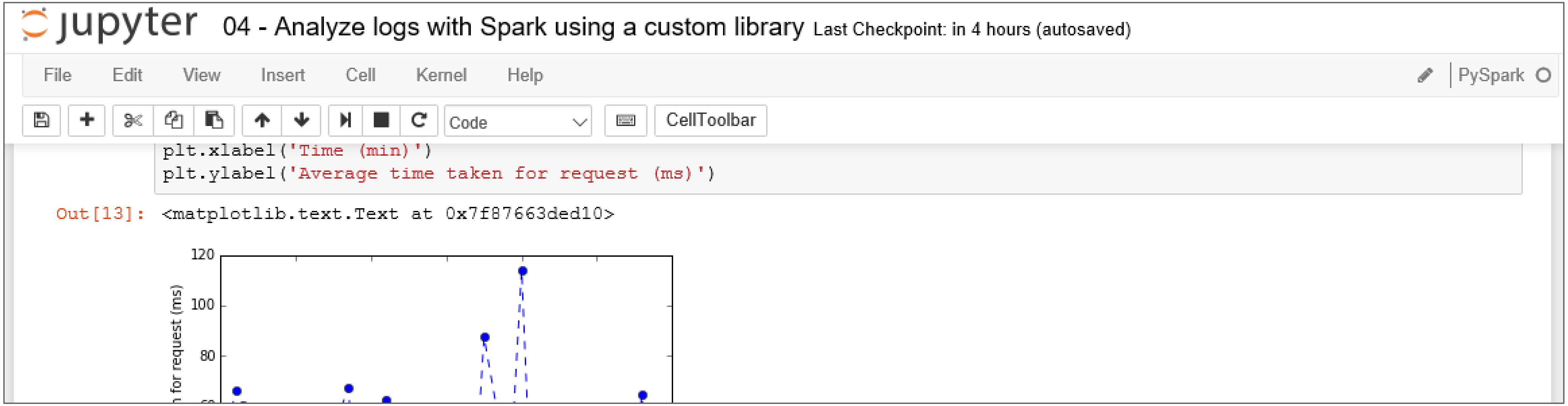
Spark Components on HDInsight

- Spark Core
  - Includes Spark SQL, Spark Streaming, GraphX, and MLlib
- Anaconda
- Livy
- Jupyter Notebooks
- ODBC Driver for connecting from BI tools (Power BI, Tableau)



Jupyter Notebooks on HDInsight

- Provide a browser-based interface for working with text, code, equations, plots, graphics, and interactive controls in a single document.
- Include preset Spark & Hive contexts (sc & sqlContext, respectively)


jupyter
04 - Analyze logs with Spark using a custom library Last Checkpoint: in 4 hours (autosaved)
File
Edit
View
Insert
Cell
Kernel
Help
PySpark
Code
CellToolbar
120
100
80
n for request (ms)

```
plt.xlabel('Time (min)')
plt.ylabel('Average time taken for request (ms)')
```

```
Out[13]: <matplotlib.text.Text at 0x7f87663ded10>
```

Microsoft Azure

Items of Note About HDInsight

- There is no “suspend” on HDInsight clusters
  - Provision the cluster, do work, then delete the cluster to avoid unnecessary charges
  - Storage can be decoupled from the cluster and reused across deployments
- Can deploy from the portal, but often scripted in practice
  - Easier/repeatable creation and deletion

Hands-On Lab

Azure HDInsight

HDInsight Spark HOL.html or HDInsight Hadoop HOL.html

Microsoft

The information herein is for informational purposes only and represents the current view of Microsoft Corporation as of the date of this presentation. Because Microsoft must respond to changing market conditions, it should not be interpreted to be a commitment on the part of Microsoft, and Microsoft cannot guarantee the accuracy of any information provided after the date of this presentation. MICROSOFT MAKES NO WARRANTIES, EXPRESS, IMPLIED OR STATUTORY, AS TO THE INFORMATION IN THIS PRESENTATION.