Choosing a data analytics technology in Azure

02/12/2018 • 4 minutes to read • Contributors 👤 🧑 👤 🧑 🧑

In this article

The goal of most big data solutions is to provide insights into the data through analysis and reporting. This can include preconfigured reports and visualizations, or interactive data exploration.

What are your options when choosing a data analytics technology?

There are several options for analysis, visualizations, and reporting in Azure, depending on your needs:

- Power BI
- Jupyter Notebooks
- Zeppelin Notebooks
- Microsoft Azure Notebooks

Power BI

Power BI is a suite of business analytics tools. It can connect to hundreds of data sources, and can be used for ad hoc analysis. See this list of the currently available data sources. Use Power BI Embedded to integrate Power BI within your own applications without requiring any additional licensing.

Organizations can use Power BI to produce reports and publish them to the organization. Everyone can create personalized dashboards, with governance and security built in. Power BI uses Azure Active Directory (Azure AD) to authenticate users who log in to the Power BI service, and uses the Power BI login credentials whenever a user attempts to access resources that require authentication.

Jupyter Notebooks

Jupyter Notebooks provide a browser-based shell that lets data scientists create *notebook* files that contain Python, Scala, or R code and markdown text, making it an effective way to collaborate by sharing and documenting code and results in a single document.

Most varieties of HDInsight clusters, such as Spark or Hadoop, come preconfigured with Jupyter notebooks for interacting with data and submitting jobs for processing. Depending on the type of HDInsight cluster you are using, one or more kernels will be provided for interpreting and running your code. For example, Spark clusters on HDInsight provide Spark-related kernels that you can select from to execute Python or Scala code using the Spark engine.

Jupyter notebooks provide a great environment for analyzing, visualizing, and processing your data prior to building more advanced visualizations with a BI/reporting tool like Power BI.

Zeppelin Notebooks

Zeppelin Notebooks are another option for a browser-based shell, similar to Jupyter in functionality. Some HDInsight clusters come preconfigured with Zeppelin notebooks. However, if you are using an HDInsight Interactive Query (Hive

LLAP) cluster, [Zeppelin](#) is currently your only choice of notebook that you can use to run interactive Hive queries. Also, if you are using a [domain-joined HDInsight cluster,](#) Zeppelin notebooks are the only type that enables you to assign different user logins to control access to notebooks and the underlying Hive tables.

Microsoft Azure Notebooks

[Azure Notebooks](#) is an online Jupyter Notebooks-based service that enables data scientists to create, run, and share Jupyter Notebooks in cloud-based libraries. Azure Notebooks provides execution environments for Python 2, Python 3, F#, and R, and provides several charting libraries for visualizing your data, such as ggplot, matplotlib, bokeh, and seaborn.

Unlike Jupyter notebooks running on an HDInsight cluster, which are connected to the cluster's default storage account, Azure Notebooks does not provide any data. You must [load data](#) in a variety of ways, such downloading data from an online source, interacting with Azure Blobs or Table Storage, connecting to a SQL database, or loading data with the Copy Wizard for Azure Data Factory.

Key benefits:

- Free service—no Azure subscription required.
- No need to install Jupyter and the supporting R or Python distributions locally—just use a browser.
- Manage your own online libraries and access them from any device.
- Share your notebooks with collaborators.

Considerations:

- You will be unable to access your notebooks when offline.
- Limited processing capabilities of the free notebook service may not be enough to train large or complex models.

Key selection criteria

To narrow the choices, start by answering these questions:

- Do you need to connect to numerous data sources, providing a centralized place to create reports for data spread throughout your domain? If so, choose an option that allows you to connect to 100s of data sources.

- Do you want to embed dynamic visualizations in an external website or application? If so, choose an option that provides embedding capabilities.

- Do you want to design your visualizations and reports while offline? If yes, choose an option with offline capabilities.

- Do you need heavy processing power to train large or complex AI models or work with very large data sets? If yes, choose an option that can connect to a big data cluster.

Capability matrix

The following tables summarize the key differences in capabilities.

General capabilities

Capability	Power BI	Jupyter Notebooks	Zeppelin Notebooks	Microsoft Azure Notebooks

Capability	Power BI	Jupyter Notebooks	Zeppelin Notebooks	Microsoft Azure Notebooks
Connect to big data cluster for advanced processing	Yes	Yes	Yes	No
Managed service	Yes	Yes [1]	Yes [1]	Yes
Connect to 100s of data sources	Yes	No	No	No
Offline capabilities	Yes [2]	No	No	No
Embedding capabilities	Yes	No	No	No
Automatic data refresh	Yes	No	No	No
Access to numerous open source packages	No	Yes [3]	Yes [3]	Yes [4]
Data transformation/cleansing options	Power Query, R	40 languages, including Python, R, Julia, and Scala	20+ interpreters, including Python, JDBC, and R	Python, F#, R
Pricing	Free for Power BI Desktop (authoring), see pricing for hosting options	Free	Free	Free
Multiuser collaboration	Yes	Yes (through sharing or with a multiuser server like JupyterHub)	Yes	Yes (through sharing)

[1] When used as part of a managed HDInsight cluster.

[2] With the use of Power BI Desktop.

[2] You can search the Maven repository for community-contributed packages.

[3] Python packages can be installed using either pip or conda. R packages can be installed from CRAN or GitHub. Packages in F# can be installed via nuget.org using the Paket dependency manager.