20230919

Succeed in the Life Sciences with R/Python and the Cloud

Colby T. Ford, Ph.D.

Tuple

Caveats and Considerations

Everything in this talk is public, but opinions are my own.

This talk references many of Microsoft services but is totally applicable to other clouds.

This content spans scientific, technical, academic, and industry viewpoints.

About Me

Colby T. Ford, Ph.D.
Computational Biologist and Cloud AI Architect

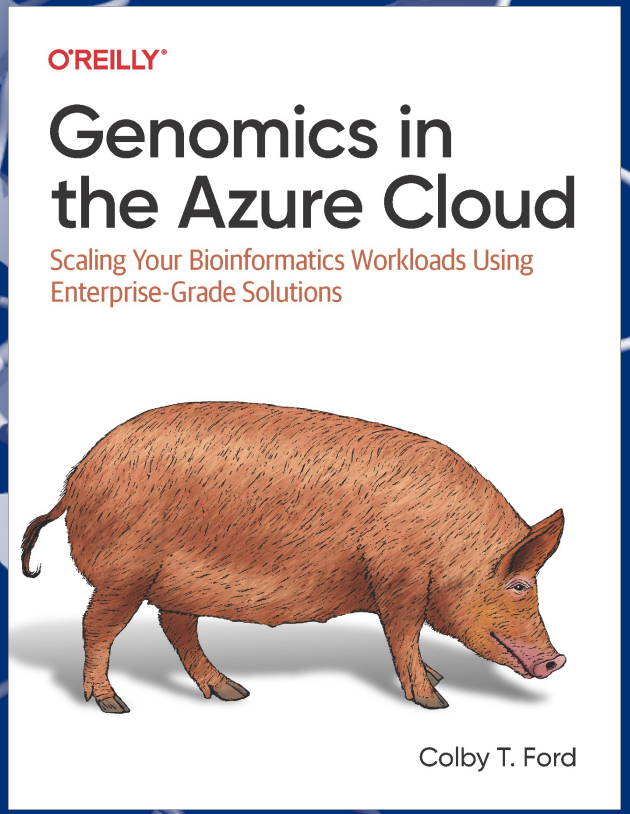
<Tuple> Founder, Principal Consultant
tuple.xyz

amissa Co-Founder, V.P. Of Technology
amissa.com

UNIVERSITY OF NORTH CAROLINA CHARLOTTE Visiting Scholar, CIPHER Research Center
cipher.charlotte.edu

Research Interests:

- Infectious diseases (SARS-CoV-2, *Plasmodium sp.*, *E. coli*)
- Human genomics (Oncology, rare diseases, etc.)
- Protein structure design (mAbs, HLA-TCR, etc.)
- Scalable cloud architectures, bioinformatics pipelines, bioAI



AzureGenomics.com

MVP Microsoft® Most Valuable Professional

Microsoft MCT

Talk Topics

Data Storage and Organization

Scaling Analyses with Cloud Compute

Case Study

* not idiot proof

Scalability

Security*

Collaboration

Beneficial Cloud Bits...

Automation

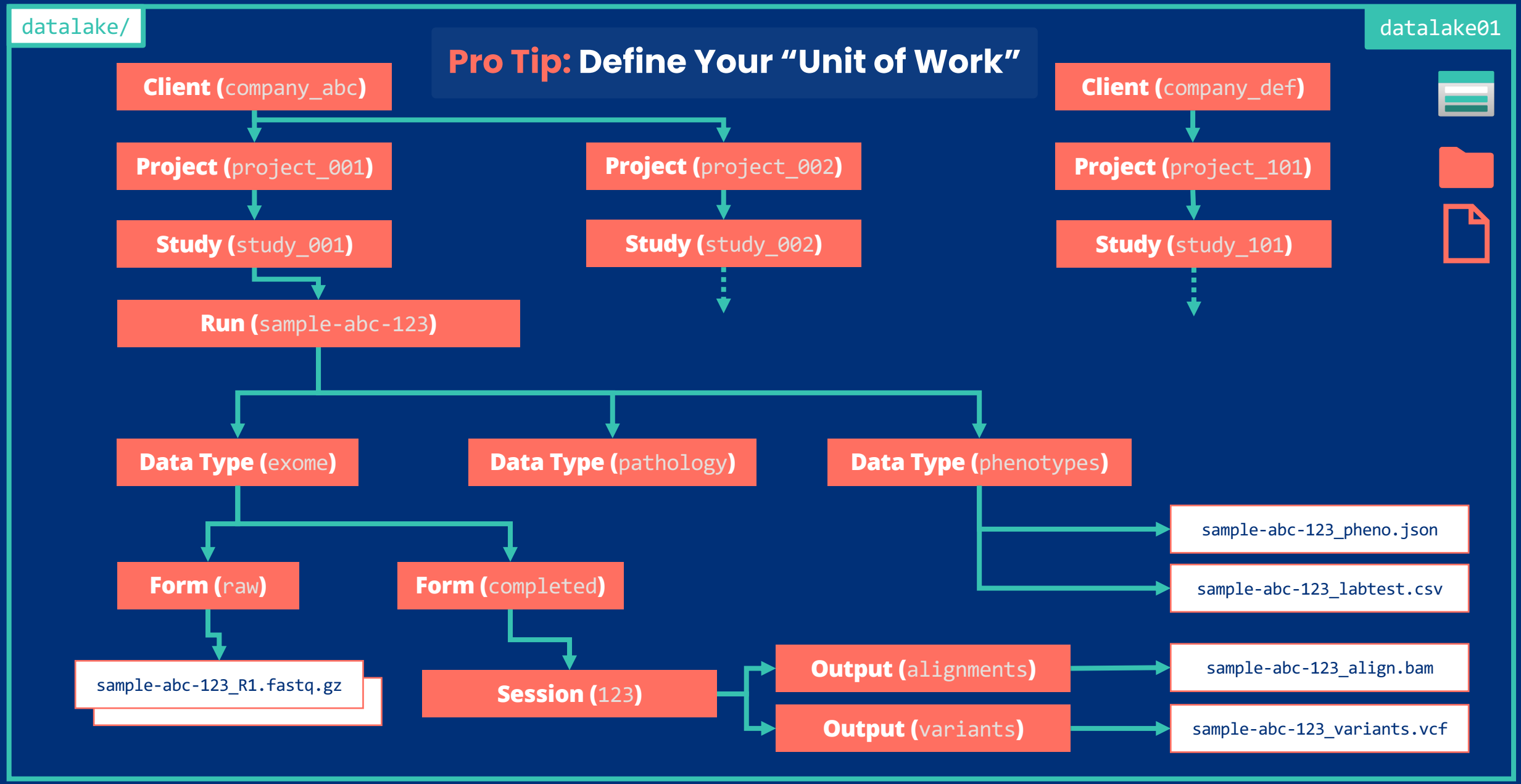
Data Storage and Organization

“The Chair”
(Base Edition)

“The Chair”
(Real Edition)

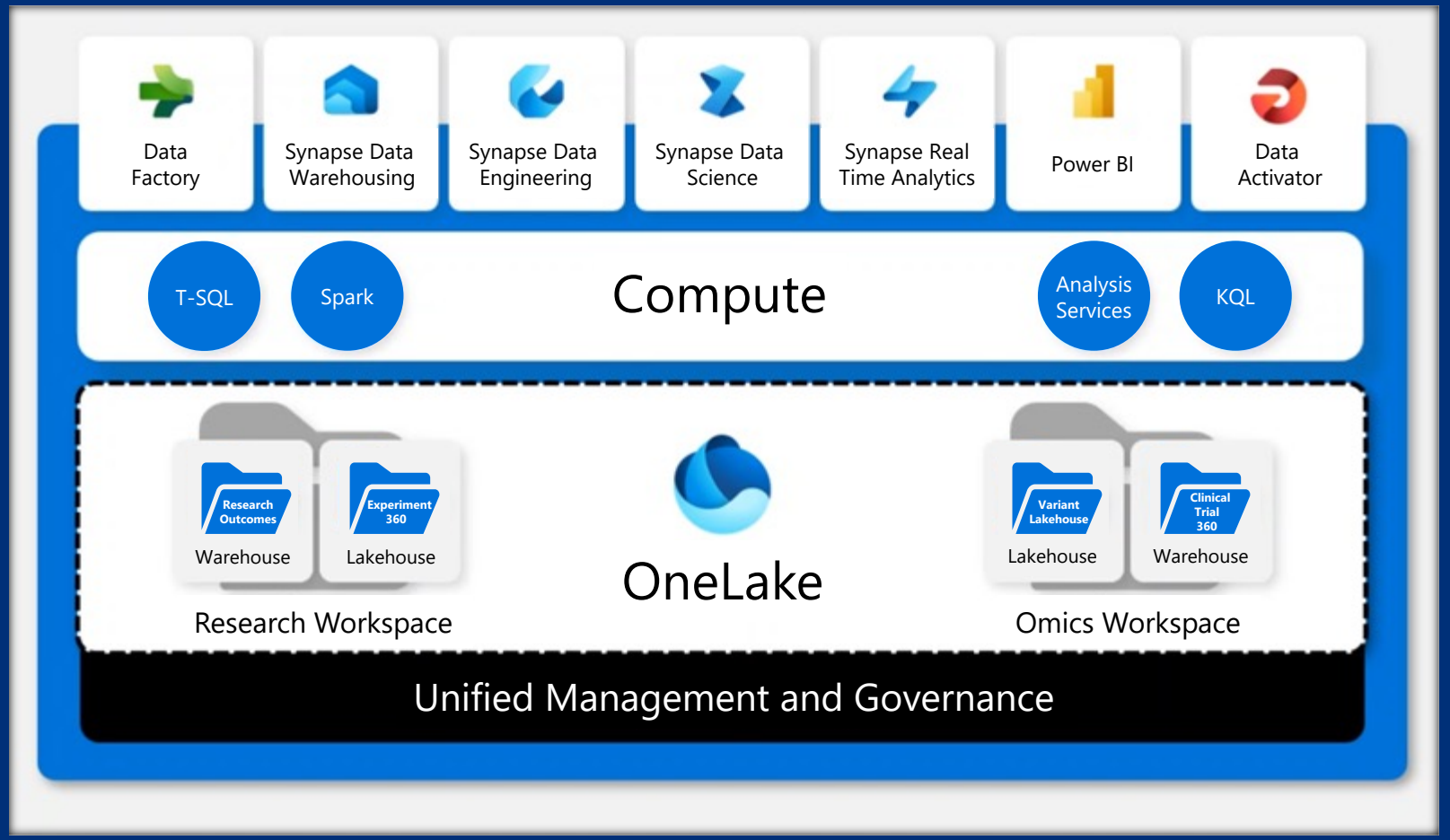
“The Chair”
(Shelf Edition)

“The Chair”
(Ideal Edition)


datalake/
datalake01
Pro Tip: Define Your "Unit of Work"
Client (company_abc)
Project (project_001)
Project (project_002)
Study (study_001)
Study (study_002)
Run (sample-abc-123)
Client (company_def)
Project (project_101)
Study (study_101)
Data Type (exome)
Data Type (pathology)
Data Type (phenotypes)
Form (raw)
Form (completed)
sample-abc-123_R1.fastq.gz
Session (123)
Output (alignments)
Output (variants)
sample-abc-123_pheno.json
sample-abc-123_labtest.csv
sample-abc-123_align.bam
sample-abc-123_variants.vcf

Data Lakes and Lakehouses

OneLake: Unification to Break Down Data Siloes



Shared Compute

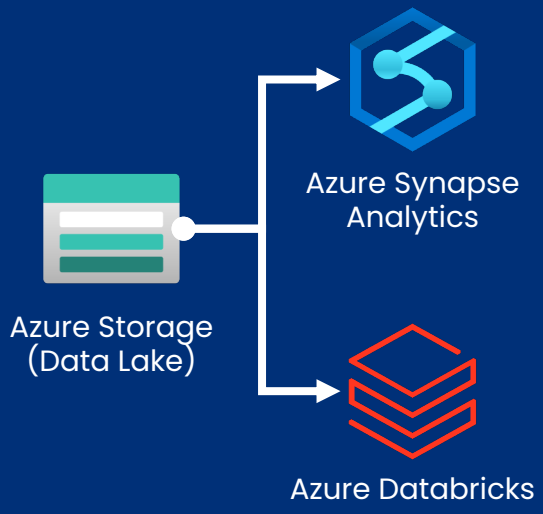
One Security

All-in on .Parquet

Why Organization Matters

Scalable Queries Across Your Data

Using tools like Azure Synapse Analytics or Azure Databricks, we can query across sets of files in a data lake (as long as it's organized).



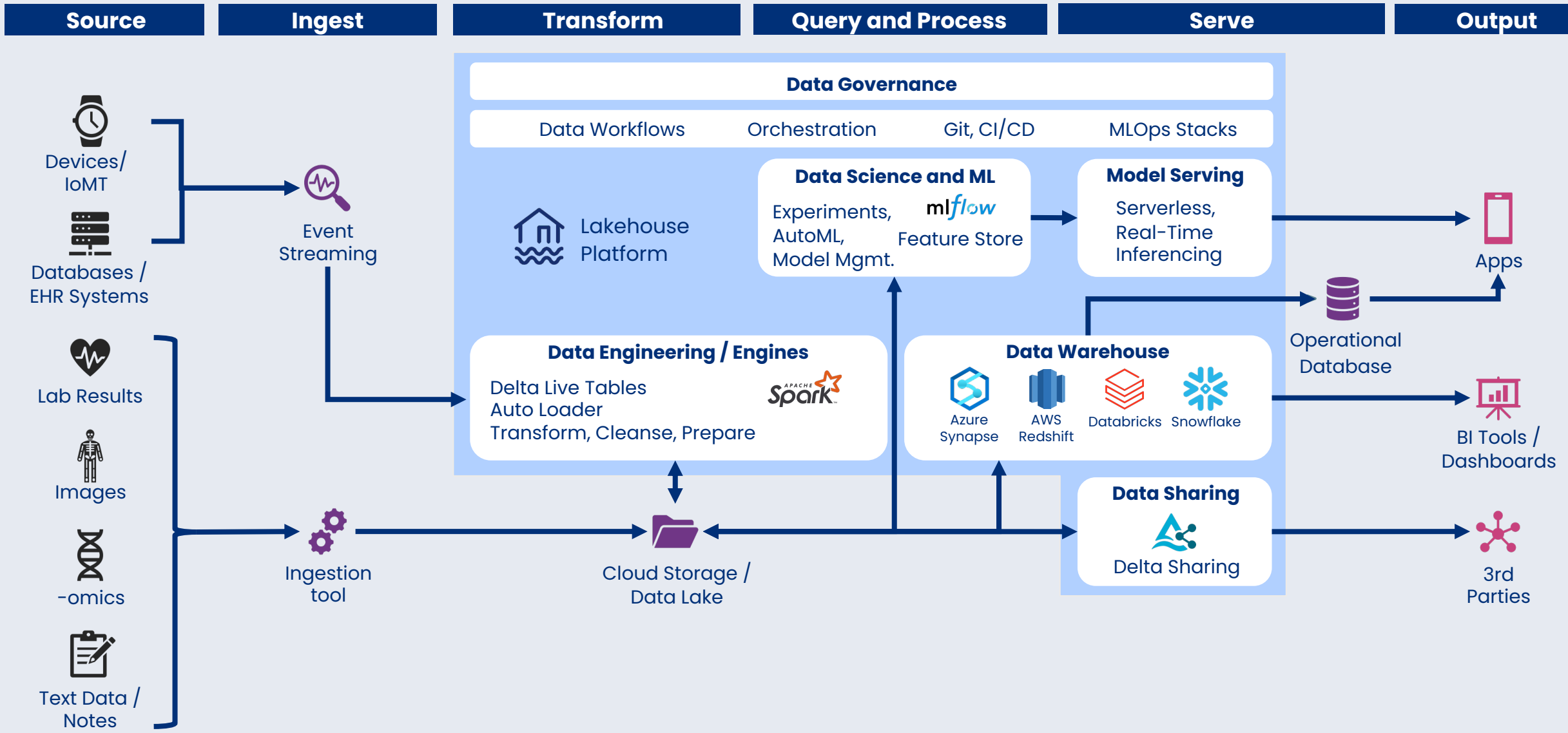
"How many of the samples in Study B have Gene *ERBB2 expression > 10 TPM* across all RNA-seq analyses?"

"Return a combined list of all spectrophotometer readings for Client Q's studies in 2022."

"What is the *average abundance ratio* for all Proteomics runs for Protein *P14678* across all of Client X's projects?"

```
datalake/project_A/study_B/rnaseq/*/completed/expression/*.csv
```

Example Architecture



Scaling Analyses with Cloud Compute

Why Cloud Compute?

Elastic Scalabilty

Orchestrate Complex Data Transformations and Analyses

Automate Repetitive Tasks

Enable Analytical Collaboration

Improve Scientific Reproducibility

Some Azure Compute Services:

Azure Synapse Analytics

Azure Kubernetes Service

Azure Databricks

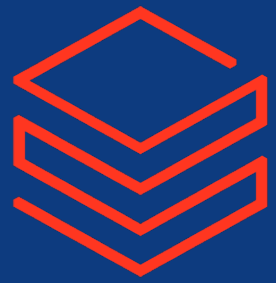
Azure Functions

Azure Machine Learning

Azure HPC Services

Running R/Python IDEs on...

Azure Machine Learning

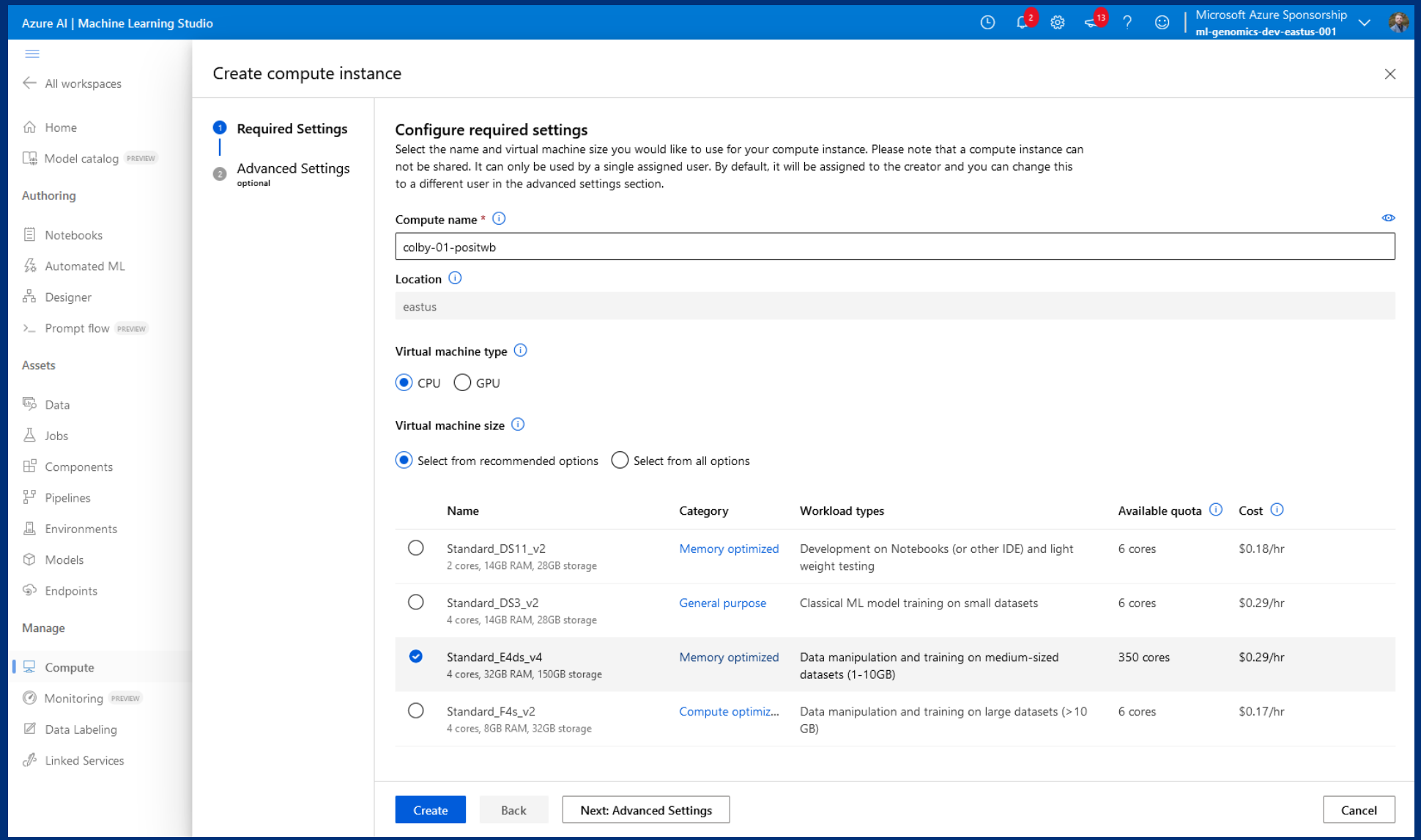


Azure Databricks

Azure Machine Learning

1

Create a Compute Instance



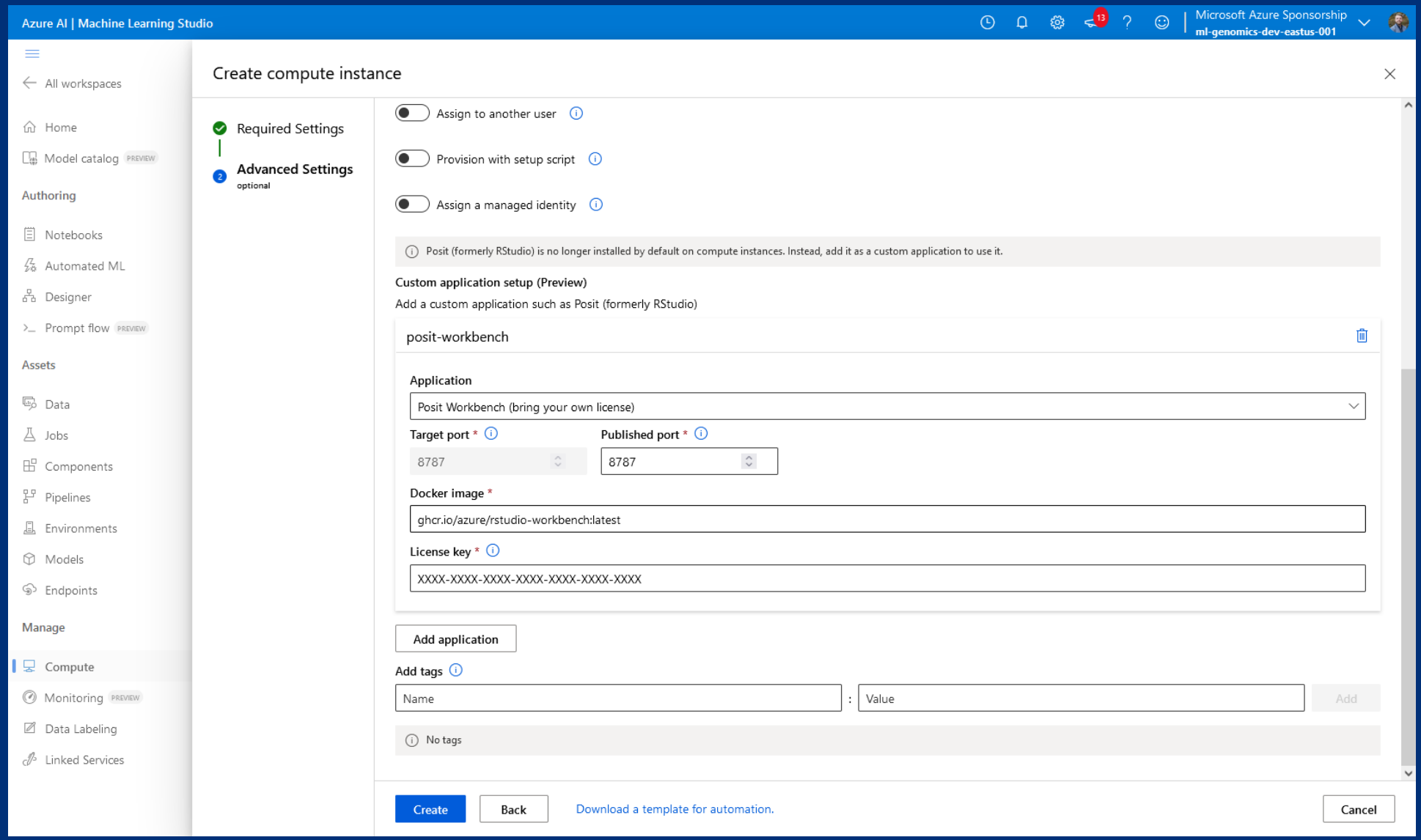
DockerHub: rstudio/rstudio-workbench

Azure Machine Learning

2A

Use a Docker Image

(in Advanced Settings)



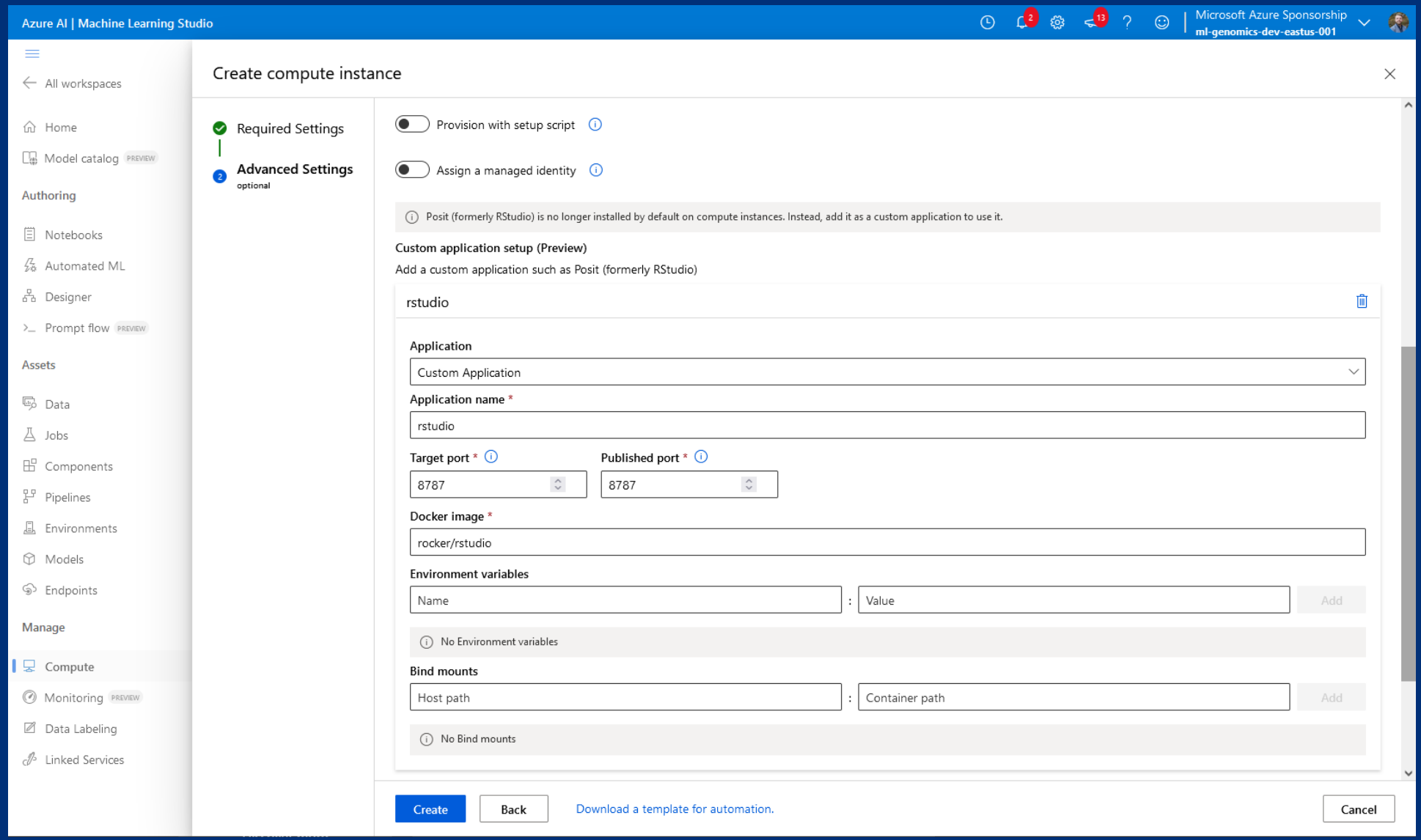
Posit Workbench BYOL

Azure Machine Learning

DockerHub: rocker/rstudio

Use a Docker Image

(in Advanced Settings)

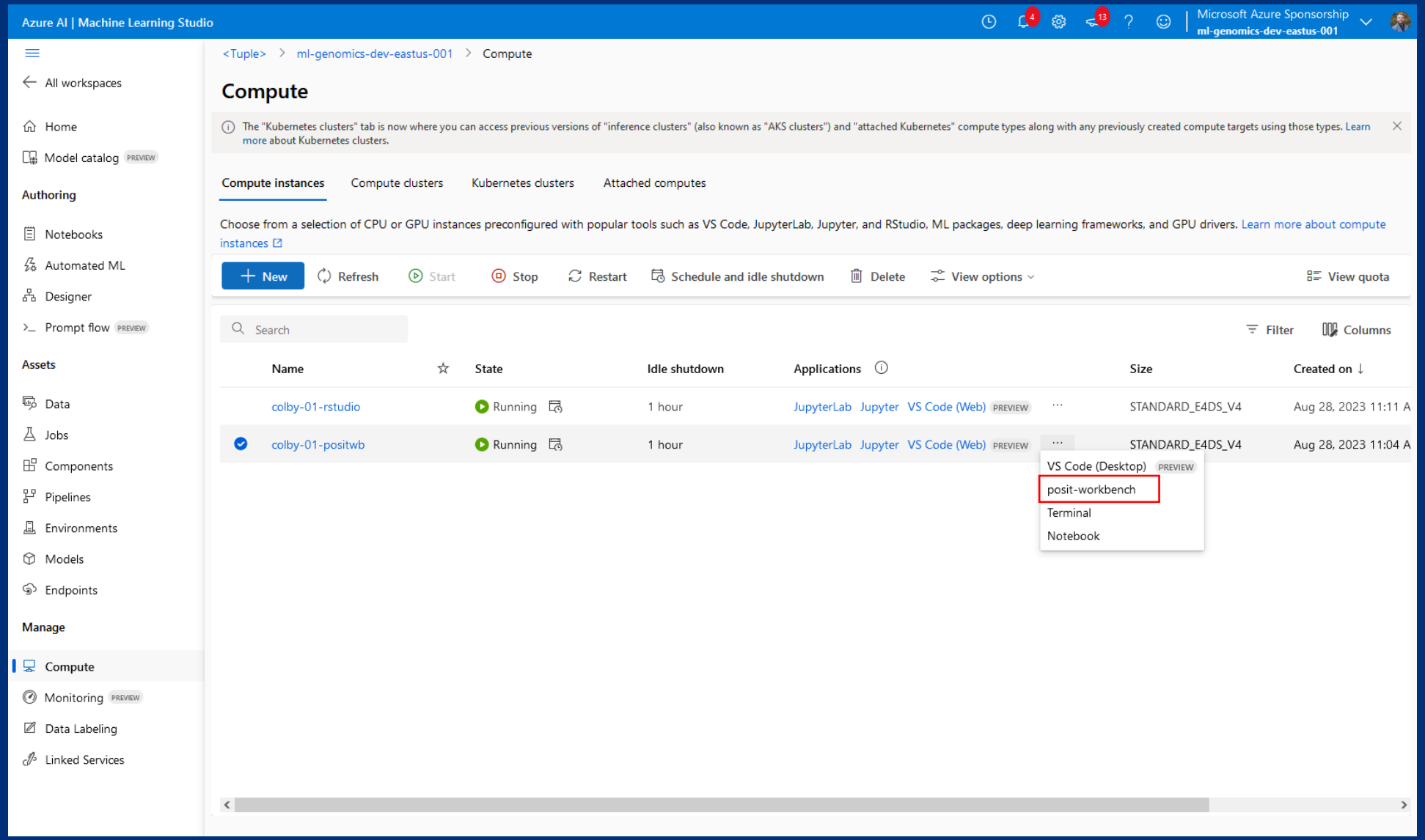


Rocker RStudio Server (Open-Source)

Azure Machine Learning

3

Open the IDE (in Applications)



Azure Machine Learning

What's Included?

as of August 2023

...in Azure ML Compute Instances

- Python 3.8 & 3.10
- R 4.3.0
- JupyterLab 3.2.4
- VSCode 1.81.1

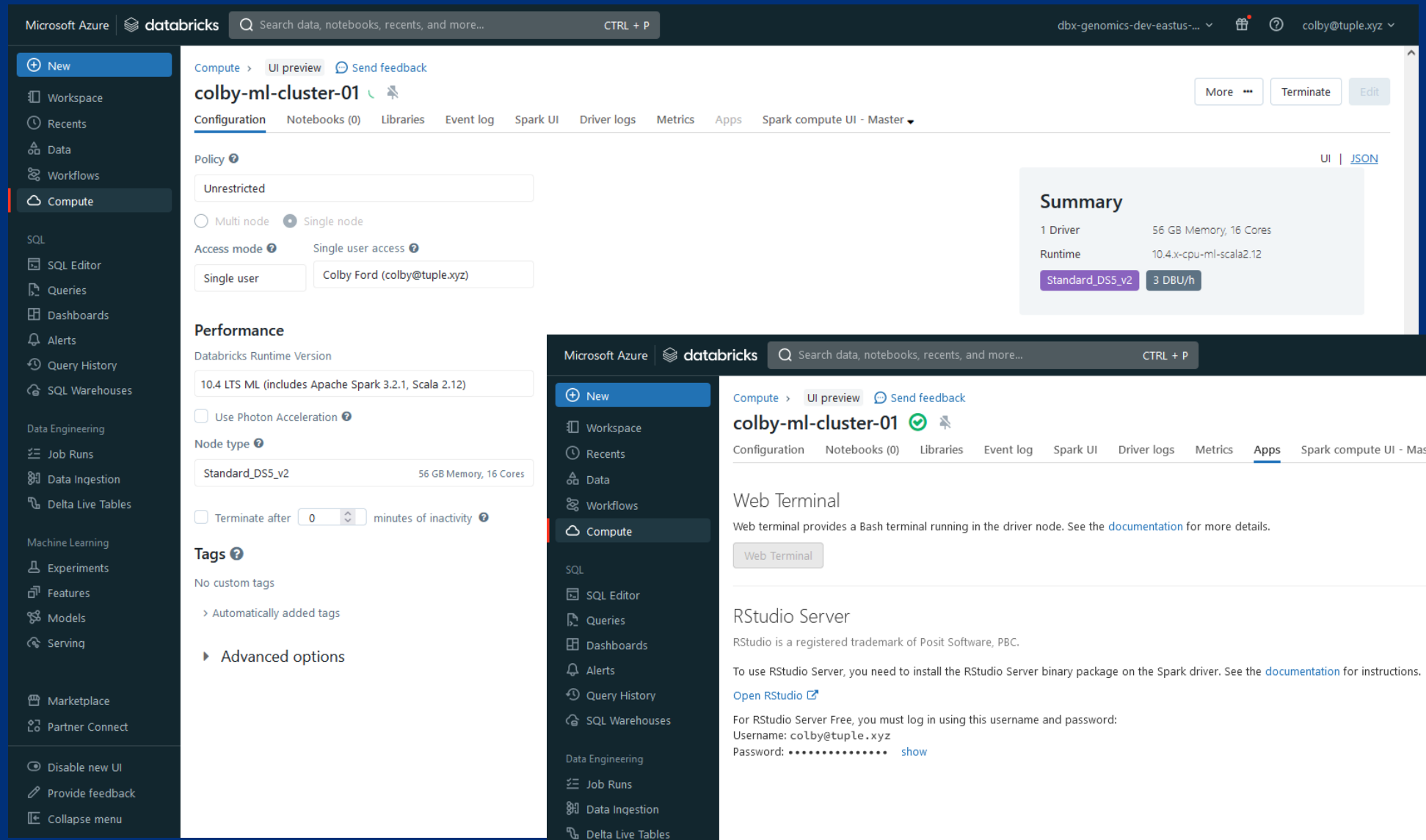
...with the 'jammy' Posit Workbench Image

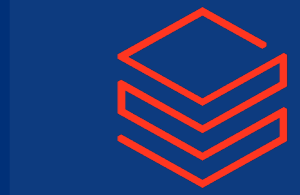
- Python 3.10.12
- R 4.2.3
- RStudio Pro 2023.06.1
- JupyterLab 3.6.5
- VSCode 4.12.0

Azure Databricks

1

Create a Cluster
(w/ ML Runtime)

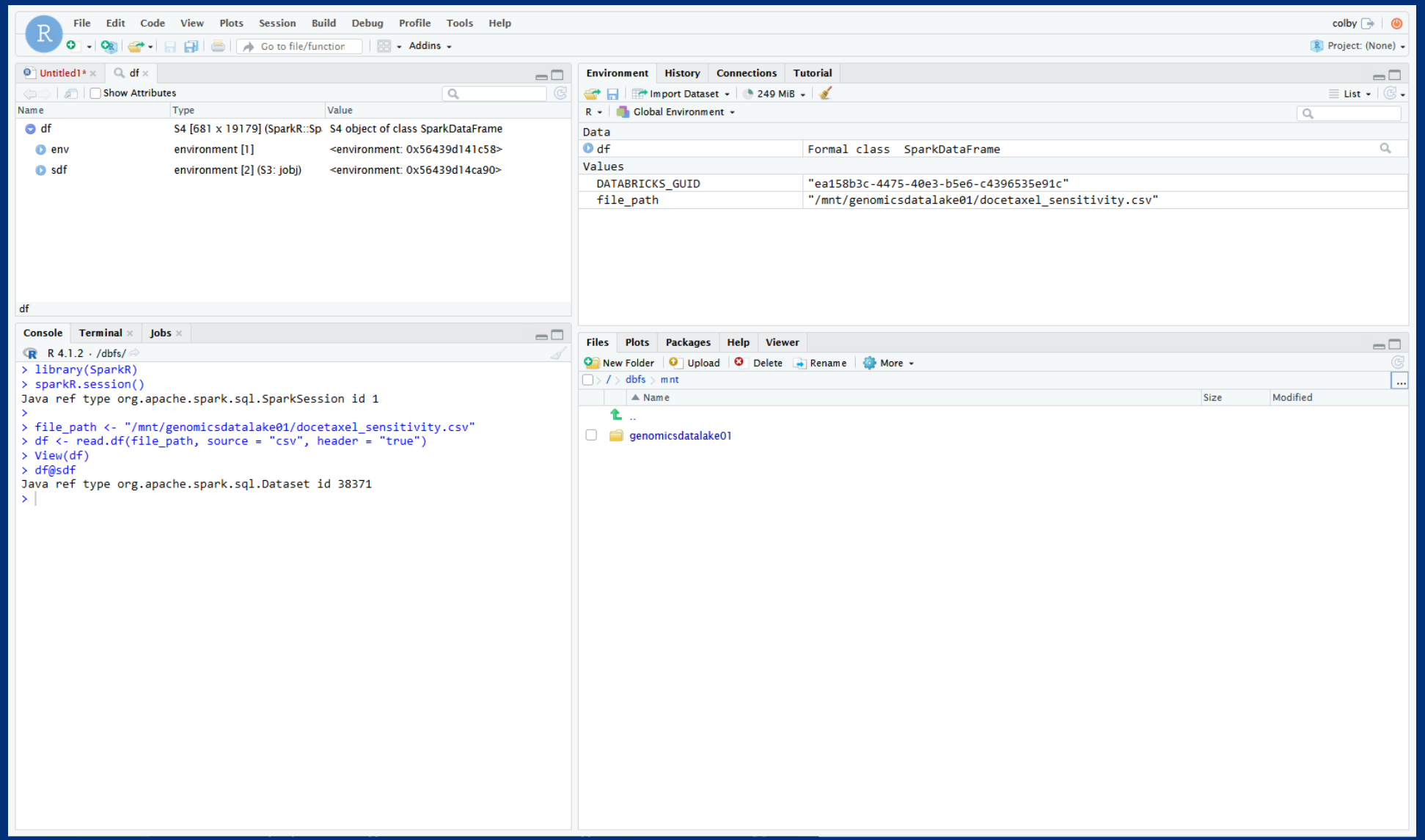




Azure Databricks

2A

Use RStudio Server



Azure Databricks

2B

Use the Databricks Notebook

Switch between R, Python, Scala, SQL, Shell, Markdown...

Microsoft Azure | databricks | Search data, notebooks, recents, and more... CTRL + P | dbx-genomics-dev-eastus-... | colby@tuple.xyz

New | Workspace | Recents | Data | Workflows | Compute

SQL: SQL Editor | Queries | Dashboards | Alerts | Query History | SQL Warehouses

Data Engineering: Job Runs | Data Ingestion | Delta Live Tables

Machine Learning: Experiments | Features | Models | Serving

Marketplace | Partner Connect | Disable new UI | Provide feedback | Collapse menu

Polyglot Example Python ☆

File Edit View Run Help Last edit was 42 minutes ago Provide feedback

Run all | colby-ml-cluster-01 | Schedule | Share

Cmd 1

Markdown

Polyglot Notebook Example

Cmd 2

Python (Default)

```
1  print("Hello from PySpark")
```

```
Hello from PySpark
```

Command took 0.72 seconds -- by colby@tuple.xyz at 8/28/2023, 1:19:22 PM on colby-ml-cluster-01

Cmd 3

Read Drug Sensitivity Data in Python

```
1  file_path = '/mnt/genomicsdatalake01/docetaxel_sensitivity.csv'
2  df = spark.read.format("csv").option("header", "true").load(file_path)
3  display(df)
```

(4) Spark Jobs

Table +

	CELL_LINE_NAME	L10_IC_50	TSPAN6	TNMD	DPM1
1	22rv1	-1.6643720768426125	2.643856189774725	0.0	6.219555769
2	2313287	-2.2657961042166934	2.985500430304885	0.0	6.778734244
3	42mgba	-2.194771295116583	4.574707046415546	0.0	6.632413641
4	5637	-2.8168508023444905	5.868637384170314	0.0	6.636044526
5	639v	-2.370915780677939	5.026800059343715	0.0	6.966130489

5 rows | Truncated data | 46.38 seconds runtime

Command took 46.38 seconds -- by colby@tuple.xyz at 8/28/2023, 1:34:45 PM on colby-ml-cluster-01

Cmd 4

R

Microsoft Azure | databricks | Search data, notebooks, recents, and more... CTRL + P

Polyglot Example Python ☆

File Edit View Run Help Last edit was 43 minutes ago Provide feedback

Cmd 4

R

```
1  %r
2  library(SparkR)
3  sparkR.session()
4  print("Hello from SparkR")
```

```
[1] "Hello from SparkR"
```

Command took 0.03 seconds -- by colby@tuple.xyz at 8/28/2023, 1:19:40 PM on colby-ml-cluster-01

Cmd 5

Read Drug Sensitivity Data in R

```
1  %r
2  file_path <- "/mnt/genomicsdatalake01/docetaxel_sensitivity.csv"
3  df <- read.df(file_path, source = "csv", header = "true")
4
5  display(df)
```

(4) Spark Jobs

Table +

	CELL_LINE_NAME	L10_IC_50	TSPAN6	TNMD	DPM1
1	22rv1	-1.6643720768426125	2.643856189774725	0.0	6.219555
2	2313287	-2.2657961042166934	2.985500430304885	0.0	6.778734
3	42mgba	-2.194771295116583	4.574707046415546	0.0	6.632413
4	5637	-2.8168508023444905	5.868637384170314	0.0	6.636044
5	639v	-2.370915780677939	5.026800059343715	0.0	6.966130

5 rows | Truncated data | 2.07 minutes runtime

Command took 2.07 minutes -- by colby@tuple.xyz at 8/28/2023, 1:53:34 PM on colby-ml-cluster-01

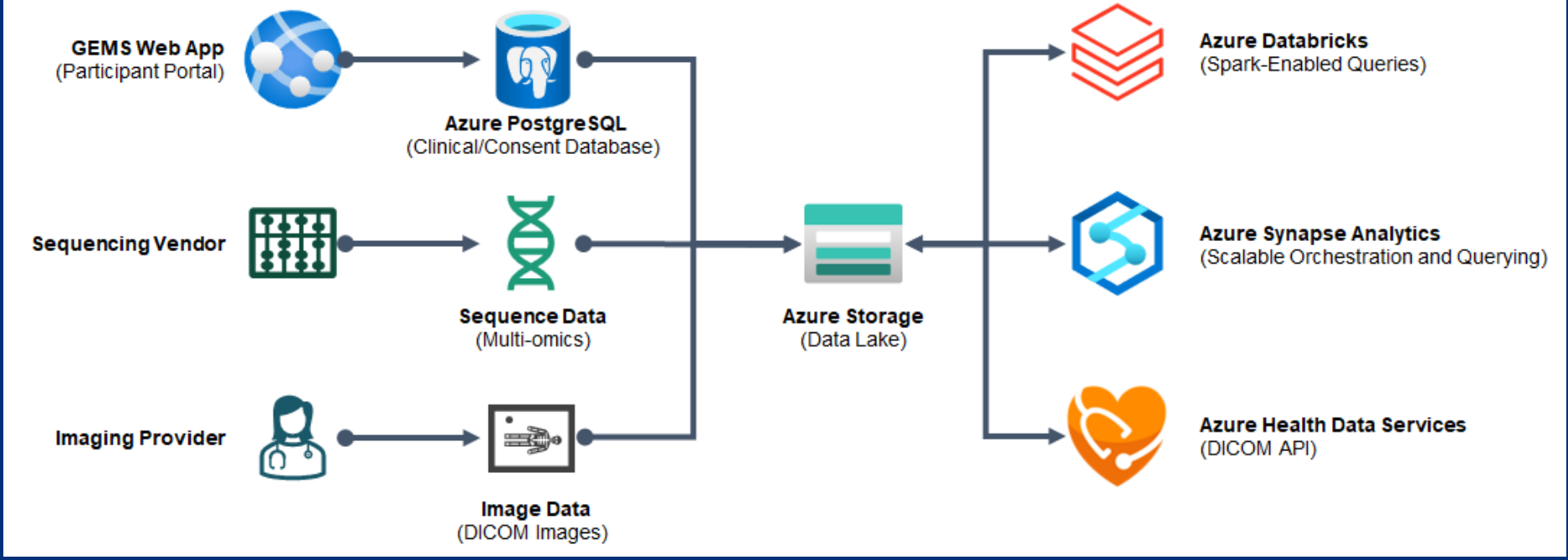
Shift+Enter to run
Shift+Ctrl+Enter to run selected text

Case Study

Fondation 101 Génomes

- Belgium-based rare disease research foundation
- First disease: Marfan syndrome

- **Solution:** Data lake-centric design with analyses housed in Databricks and Synapse



Fondation 101 Génomes

18 Seconds	10 Seconds	17 Minutes

WGS Samples:
49 Marfan +
112 Control

→ Read in variant (VCF) data

→ Filter to gene / region of interest

→ Pivot by variant and report zygosity

	sampleId	chr[redacted]:disruptive_inframe_insertion	chr[redacted]:intron_variant
1	[redacted]	wildtype	wildtype
2		homozygous	wildtype
3		wildtype	wildtype
4		wildtype	wildtype
5		wildtype	wildtype
6		wildtype	homozygous
7		wildtype	wildtype
8		wildtype	wildtype

In Summary...

Running workloads in the cloud offers scalability, flexibility, and automation.

Data lakes offer a single source for your data, but it needs to be organized to be useful.

Connected compute services allow for you to retrieve data from your sources in a similar way to working locally.

Connect with Me:

colby@tuple.xyz

@colbyford

Questions?

<Tuple>