



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

CHARLOTTE
UNIVERSITY OF NORTH CAROLINA

Visiting Scholar, CIPHER Research Center
cipher.charlotte.edu

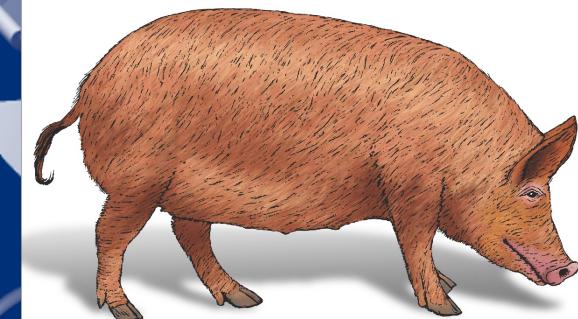
Research Interests:

- 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

O'REILLY®

Genomics in the Azure Cloud

Scaling Your Bioinformatics Workloads Using Enterprise-Grade Solutions



Colby T. Ford

AzureGenomics.com



Microsoft®
Most Valuable
Professional



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”

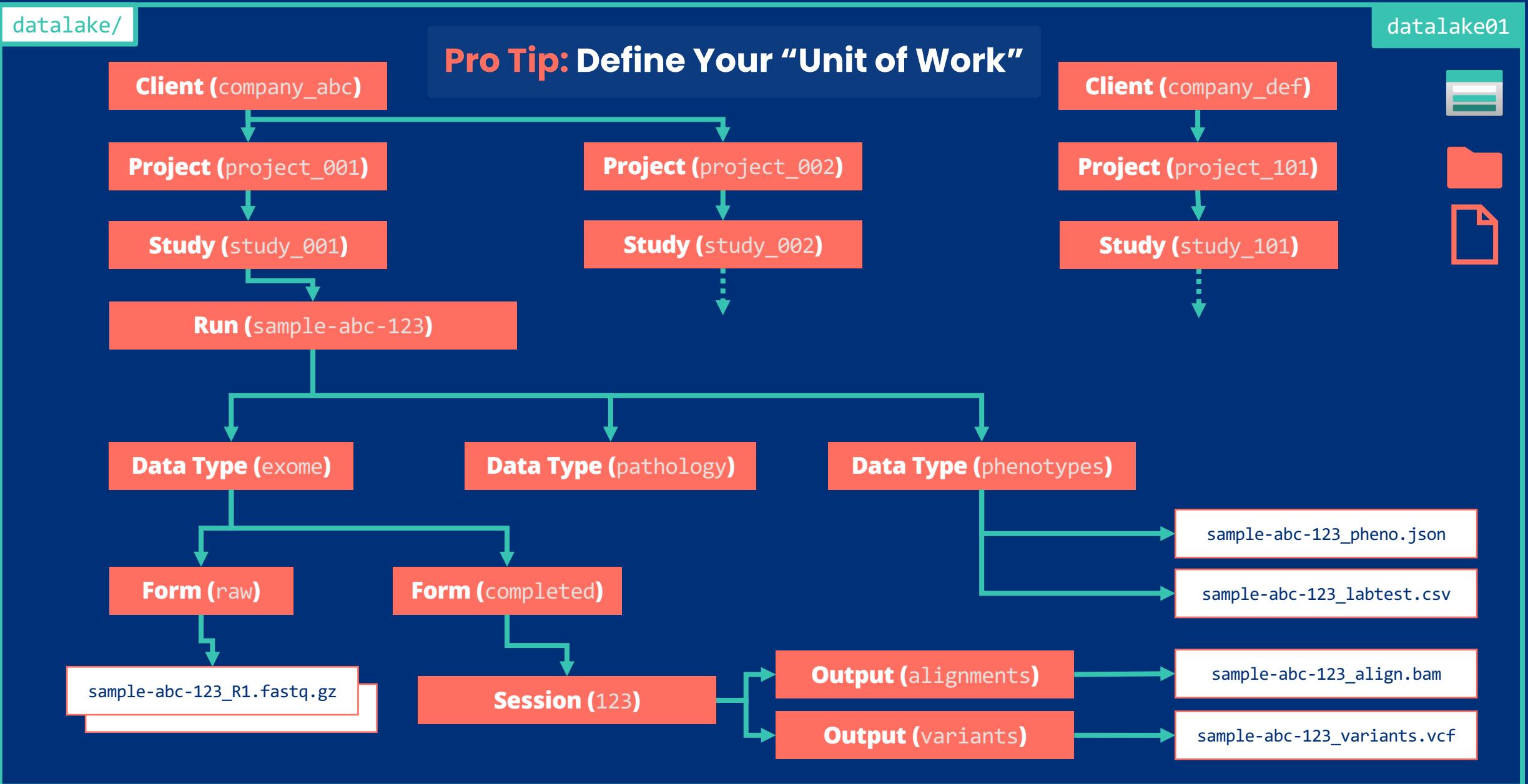
(Shelf Edition)



“The Chair”

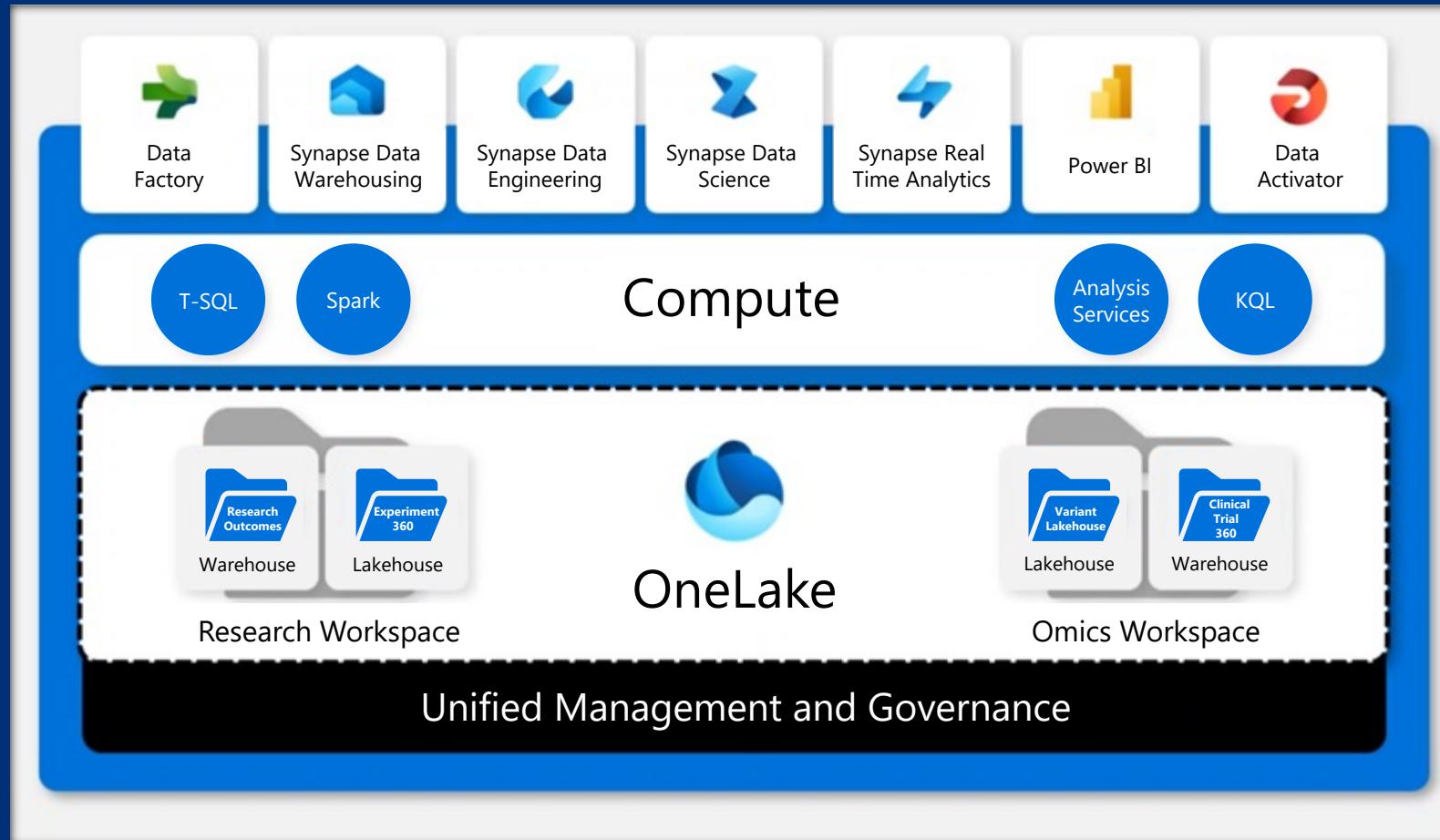
(Ideal Edition)





Data Lakes and Lakehouses

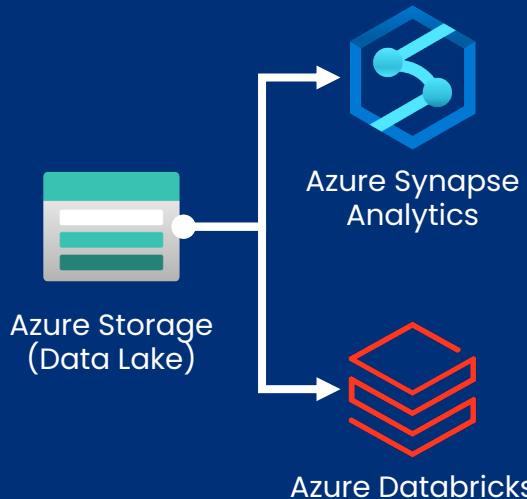
OneLake: Unification to Break Down Data Siloes



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).



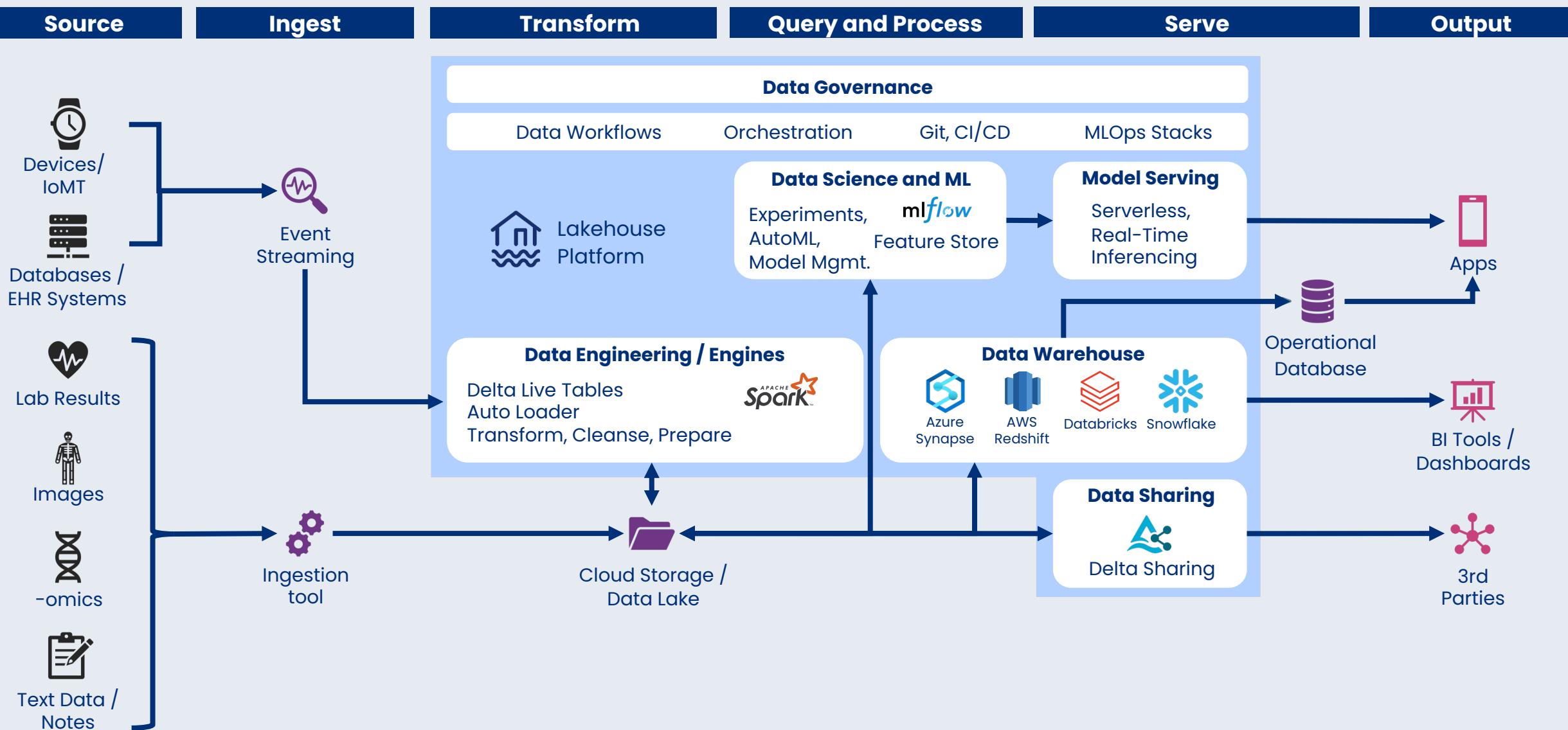
"How many of the samples in Study B have Gene *ERBB2* expression > 10 TPM across all RNA-seq analyses?"

"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 Scalability

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

Azure AI | Machine Learning Studio

All workspaces

Home

Model catalog PREVIEW

Authoring

Notebooks

Automated ML

Designer

Prompt flow PREVIEW

Assets

Data

Jobs

Components

Pipelines

Environments

Models

Endpoints

Manage

Compute

Monitoring PREVIEW

Data Labeling

Linked Services

Create compute instance

Required Settings

Advanced Settings optional

Configure required settings

Select the name and virtual machine size you would like to use for your compute instance. Please note that a compute instance can not be shared. It can only be used by a single assigned user. By default, it will be assigned to the creator and you can change this to a different user in the advanced settings section.

Compute name *

colby-01-positwb

Location *

eastus

Virtual machine type *

CPU GPU

Virtual machine size *

Select from recommended options Select from all options

Name	Category	Workload types	Available quota *	Cost *
Standard_DS11_v2 2 cores, 14GB RAM, 28GB storage	Memory optimized	Development on Notebooks (or other IDE) and light weight testing	6 cores	\$0.18/hr
Standard_DS3_v2 4 cores, 14GB RAM, 28GB storage	General purpose	Classical ML model training on small datasets	6 cores	\$0.29/hr
Standard_E4ds_v4 4 cores, 32GB RAM, 150GB storage	Memory optimized	Data manipulation and training on medium-sized datasets (1-10GB)	350 cores	\$0.29/hr
Standard_F4s_v2 4 cores, 8GB RAM, 32GB storage	Compute optimiz...	Data manipulation and training on large datasets (>10 GB)	6 cores	\$0.17/hr

Create Back Next: Advanced Settings Cancel



Use a Docker Image (in Advanced Settings)

Azure AI | Machine Learning Studio

Create compute instance

Required Settings

Advanced Settings optional

Assign to another user

Provision with setup script

Assign a managed identity

Posit (formerly RStudio) is no longer installed by default on compute instances. Instead, add it as a custom application to use it.

Custom application setup (Preview)

Add a custom application such as Posit (formerly RStudio)

posit-workbench

Application

Posit Workbench (bring your own license)

Target port * 8787

Published port * 8787

Docker image *

ghcr.io/azure/rstudio-workbench:latest

License key *

XXXX-XXXX-XXXX-XXXX-XXXX-XXXX-XXXX

Add application

Add tags

Name : Value

No tags

Create Back Download a template for automation. Cancel

DockerHub: rstudio/rstudio-workbench

Posit Workbench BYOL



Azure Machine Learning

2B

Use a Docker Image (in Advanced Settings)

Azure AI | Machine Learning Studio

Create compute instance

Required Settings

Provision with setup script [\(i\)](#)

Assign a managed identity [\(i\)](#)

Posit (formerly RStudio) is no longer installed by default on compute instances. Instead, add it as a custom application to use it.

Custom application setup (Preview)

Add a custom application such as Posit (formerly RStudio)

rstudio

Application

Custom Application

Application name *

rstudio

Target port * [\(i\)](#) **Published port * [\(i\)](#)**

8787 8787

Docker image *

rocker/rstudio

Environment variables

Name	:	Value	Add
------	---	-------	---------------------

No Environment variables

Bind mounts

Host path	:	Container path	Add
-----------	---	----------------	---------------------

No Bind mounts

Create **Back** Download a template for automation. **Cancel**

Rocker RStudio Server (Open-Source)



Azure Machine Learning

3

Open
the IDE
(in Applications)

Azure AI | Machine Learning Studio

< Tuple > > ml-genomics-dev-eastus-001 > Compute

Compute

The "Kubernetes clusters" tab is now where you can access previous versions of "inference clusters" (also known as "AKS clusters") and "attached Kubernetes" compute types along with any previously created compute targets using those types. [Learn more about Kubernetes clusters.](#)

Compute instances Compute clusters Kubernetes clusters Attached computes

Choose from a selection of CPU or GPU instances preconfigured with popular tools such as VS Code, JupyterLab, Jupyter, and RStudio, ML packages, deep learning frameworks, and GPU drivers. [Learn more about compute instances](#)

+ New Refresh Start Stop Restart Schedule and idle shutdown Delete View options View quota

Search Filter Columns

Name	State	Idle shutdown	Applications	Size	Created on
colby-01-rstudio	Running	1 hour	JupyterLab Jupyter VS Code (Web)	STANDARD_E4DS_V4	Aug 28, 2023 11:11 A
colby-01-positwb	Running	1 hour	JupyterLab Jupyter VS Code (Web)	STANDARD_E4DS_V4	Aug 28, 2023 11:04 A

VS Code (Desktop) PREVIEW
posit-workbench
Terminal
Notebook

All workspaces Home Model catalog PREVIEW Authoring Notebooks Automated ML Designer Prompt flow PREVIEW Assets Data Jobs Components Pipelines Environments Models Endpoints Manage Compute Monitoring PREVIEW Data Labeling Linked Services



Azure Machine
Learning

What's Included?

as of August 2023

...in Azure ML Compute Instances



Python
3.8 & 3.10



JupyterLab
3.2.4

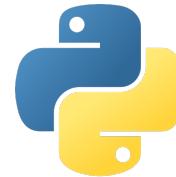


R
4.3.0



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)

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

Compute > UI preview Send feedback

colby-ml-cluster-01

Configuration Notebooks (0) Libraries Event log Spark UI Driver logs Metrics Apps Spark compute UI - Master ▾

More **Terminate** **Edit**

Policy Unrestricted

Multi node Single node

Access mode Single user access

Single user Colby Ford (colby@tuple.xyz)

UI | **JSON**

Summary

1 Driver	56 GB Memory, 16 Cores
Runtime	10.4.x-cpu-ml-scala2.12
Standard_DS5_v2	3 DBU/h

Performance

Databricks Runtime Version: 10.4 LTS ML (includes Apache Spark 3.2.1, Scala 2.12)

Use Photon Acceleration

Node type Standard_DS5_v2 56 GB Memory, 16 Cores

Terminate after 0 minutes of inactivity

Tags

No custom tags

> Automatically added tags

> Advanced options

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

Compute > UI preview Send feedback

colby-ml-cluster-01

Configuration Notebooks (0) Libraries Event log Spark UI Driver logs Metrics Apps Spark compute UI - Master ▾

Web Terminal

Web terminal provides a Bash terminal running in the driver node. See the [documentation](#) for more details.

RStudio Server

RStudio is a registered trademark of Posit Software, PBC.

To use RStudio Server, you need to install the RStudio Server binary package on the Spark driver. See the [documentation](#) for instructions.

[Open RStudio](#)

For RStudio Server Free, you must log in using this username and password:

Username: colby@tuple.xyz

Password: ***** [show](#)

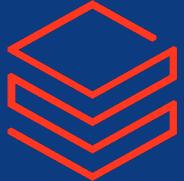
Compute **UI preview** **Send feedback**

colby-ml-cluster-01

Configuration Notebooks (0) Libraries Event log Spark UI Driver logs Metrics Apps Spark compute UI - Master ▾

More **Terminate** **Edit**

UI | **JSON**



Azure Databricks

2A

Use RStudio Server

The screenshot shows the RStudio Server interface with the following components:

- Environment View:** Shows variables `df`, `env`, and `sdf`. `df` is described as a `S4 [681 x 19179] (SparkR::Sp)` object of class `SparkDataFrame`.
- Data View:** Shows formal class `SparkDataFrame` for `df`. It also lists `DATABRICKS_GUID` and `file_path`.
- Console View:** Displays R code running in a SparkR session. The code reads a CSV file from a Databricks File System (dbfs) location.
- File Browser:** Shows a directory structure under `/dbfs/mnt/genomicsdatalake01`, containing a folder named `genomicsdatalake01`.

```
R 4.1.2 · /dbfs/ 
> library(SparkR)
> sparkR.session()
Java ref type org.apache.spark.sql.SparkSession id 1
>
> file_path <- "/mnt/genomicsdatalake01/docetaxel_sensitivity.csv"
> df <- read.df(file_path, source = "csv", header = "true")
> View(df)
> df@sdf
Java ref type org.apache.spark.sql.Dataset id 38371
>
```



Use the Databricks Notebook

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

Azure Databricks



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

Polyglot Example Python Provide feedback

File Edit View Run Help Last edit was 42 minutes ago

Run all colby-ml-cluster-01 Schedule Share

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 Notebook Example

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

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

CELL_LINE_NAME	L10_IC_50	TSPANG	TNMD	DPM1
1 22rv1	-1.6643720768426125	2.643856189774725	0.0	6.2195557691
2 2313287	-2.2657961042166934	2.985500430304885	0.0	6.7787342441
3 42mgba	-2.194771295116583	4.574707046415546	0.0	6.6324136411
4 5637	-2.8168508023444905	5.868637384170314	0.0	6.6360445260
5 639v	-2.370915780677939	5.026800059343715	0.0	6.9661304899

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

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

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 display(df)
```

(4) Spark Jobs

CELL_LINE_NAME	L10_IC_50	TSPANG	TNMD	DPM1
1 22rv1	-1.6643720768426125	2.643856189774725	0.0	6.2195557691
2 2313287	-2.2657961042166934	2.985500430304885	0.0	6.7787342441
3 42mgba	-2.194771295116583	4.574707046415546	0.0	6.6324136411
4 5637	-2.8168508023444905	5.868637384170314	0.0	6.6360445260
5 639v	-2.370915780677939	5.026800059343715	0.0	6.9661304899

5 rows | Truncated data | 2.07 minutes runtime

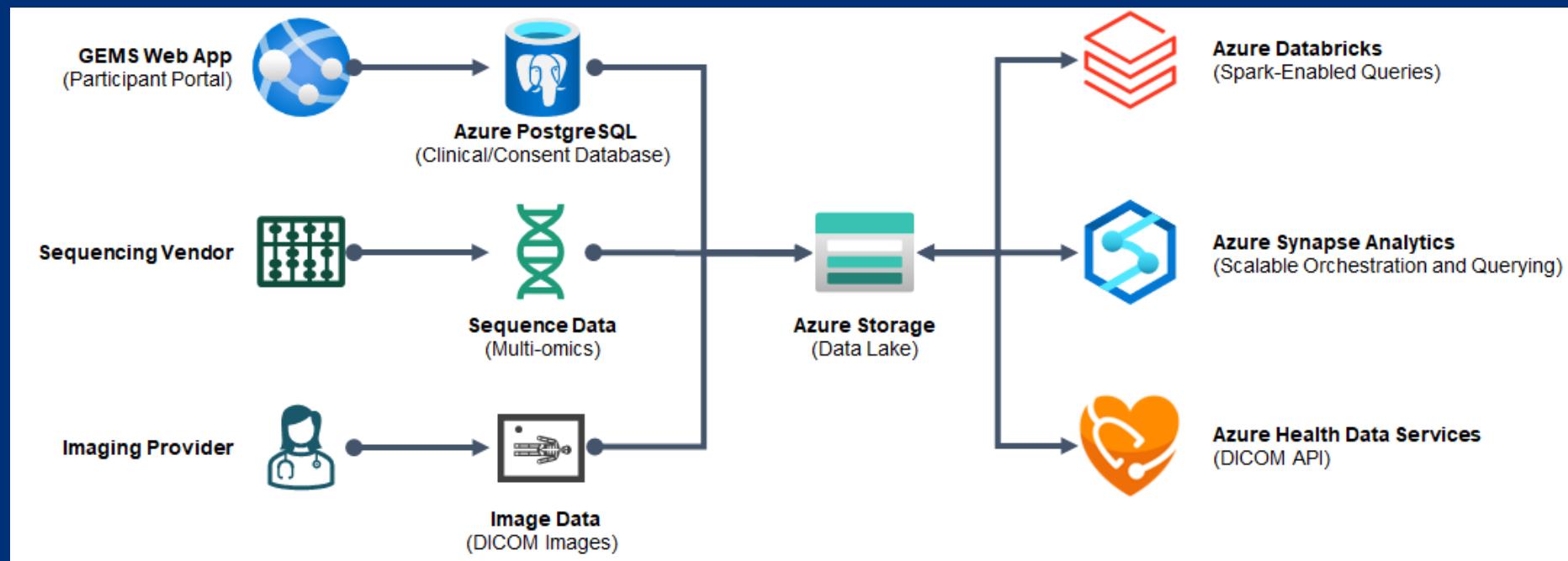
Command took 2.07 minutes -- by colby@tuple.xyz at 8/28/2023, 1:53:34 PM on colby-ml-cluster-01

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

WGS Samples:
49 Marfan +
112 Control

10 Seconds

Read in variant
(VCF) data

17 Minutes

Filter to gene /
region of
interest

Pivot by variant
and report
zygosity

	sampleId	chr	:disruptive_inframe_insertion	chr	:intron_variant
1		wildtype		wildtype	
2		homozygous		wildtype	
3		wildtype		wildtype	
4		wildtype		wildtype	
5		wildtype		wildtype	
6		wildtype		wildtype	
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
/github/    @colbyford

Questions?

<Tuple>