Scaling Biotech Workloads in 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

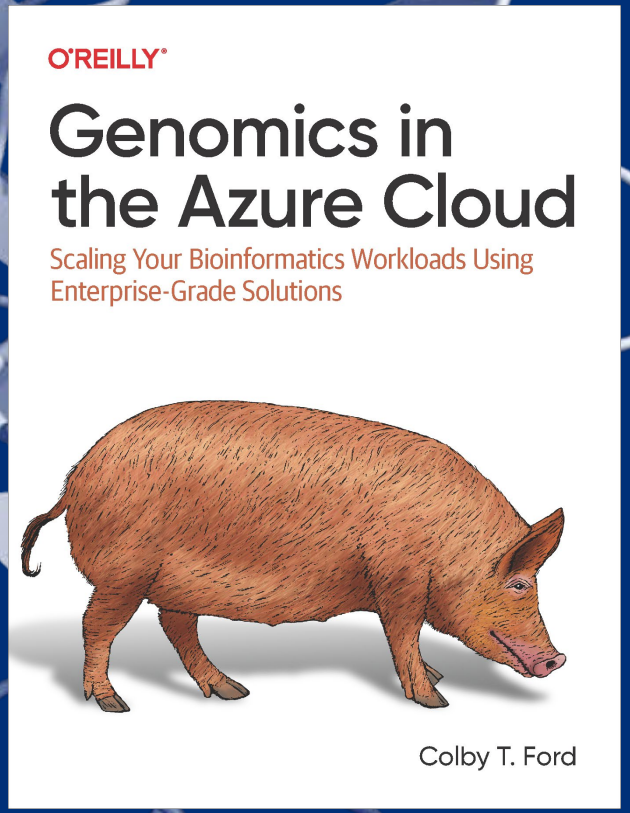
Founder, Principal Consultant
tuple.xyz

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

Associate Faculty, School of Data Science
sds.uncc.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

Talk Topics

Considerations for/against the Cloud

Data Storage and Organization

Scaling Analyses with Cloud Compute

Considerations For/Against the Cloud

Budgeting in Grants

Data Storage and Organization

Data Lakes and Lakehouses

OneLake: Unification to Break Down Data Siloes

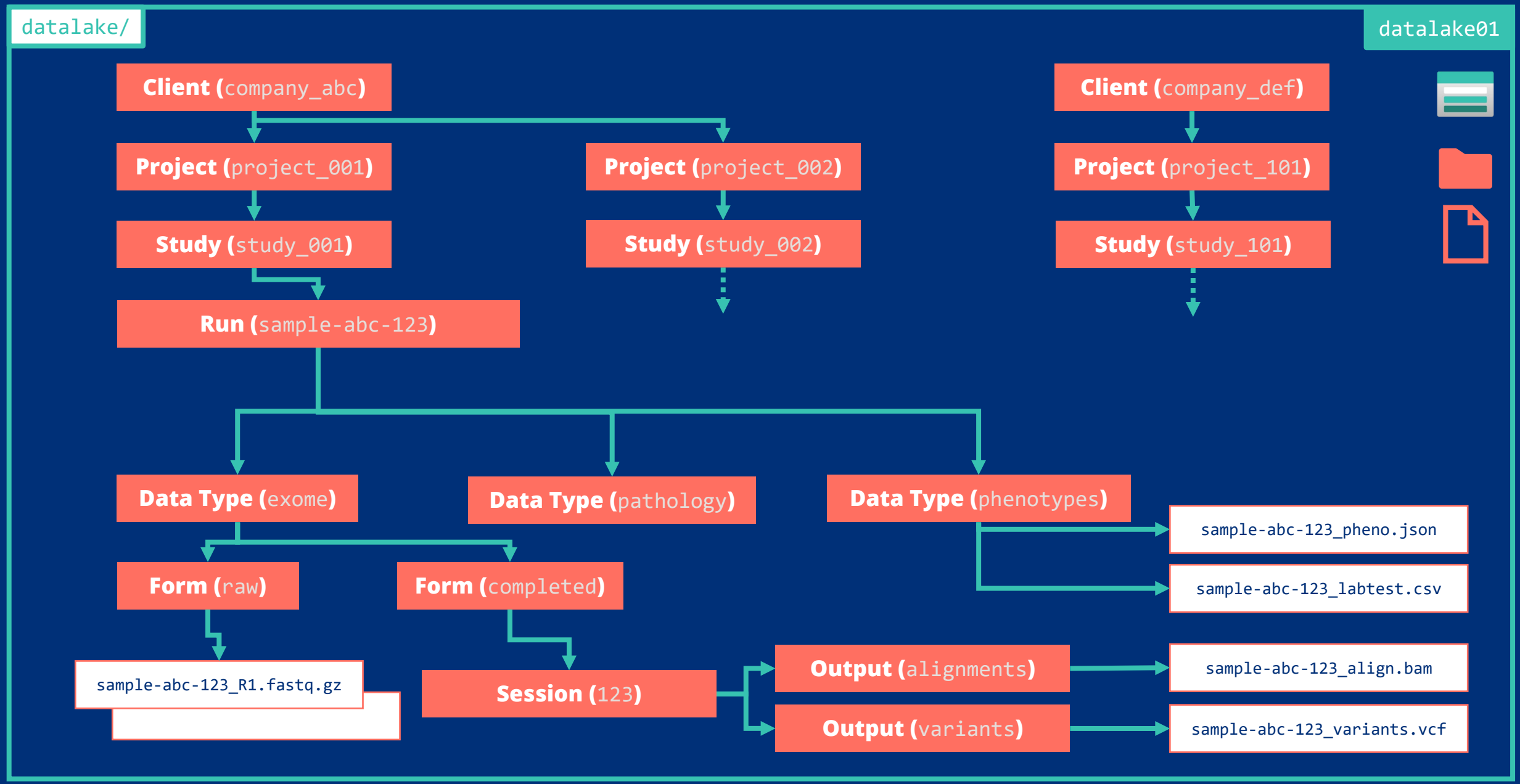
Shared Compute

One Security

All-in on .Parquet

Pro Tip:

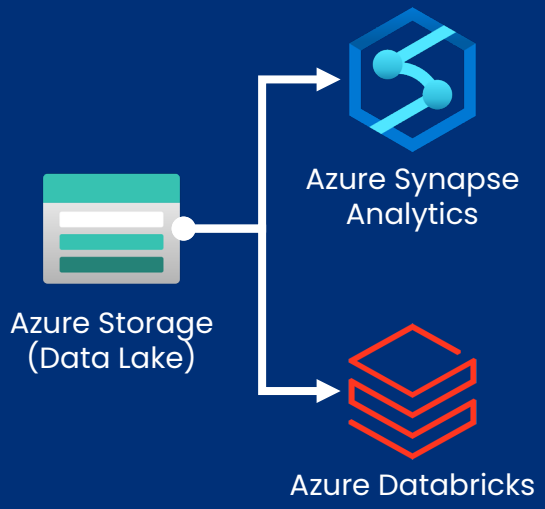
Define Your "Unit of Work"


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

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 *ABCB1 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?"

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

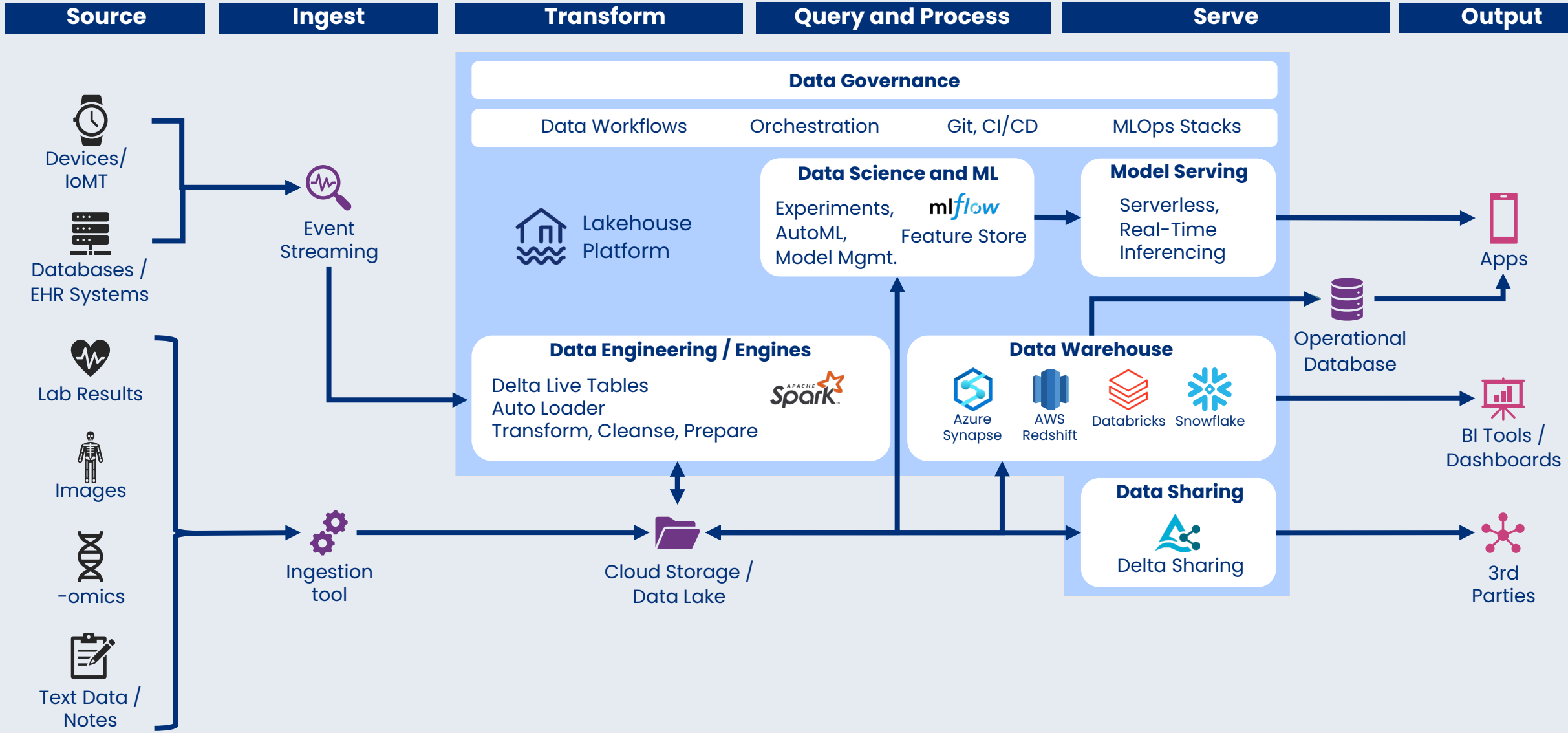
Azure Databricks

Azure Functions

Azure Machine Learning

Azure HPC Services

Example Architecture



Connect with Me:

colby@tuple.xyz

@colbyford

Questions?

<Tuple>