# Scaling Genomic Analyses in Azure

**Presented By:**
Eric Wozniak
Colby Ford, Ph.D.

BLUEGRANITE
DATA • INSIGHTS • ANALYTICS

## **Eric Wozniak** | VP – Industry, Healthcare

- Industry lead for the Healthcare vertical
  - Payors, Providers, Pharmaceuticals, Life Sciences
- Manages strategic accounts by developing go-to-market strategies and offers and building partner relationships.

## **Colby Ford** | AI Architect

- Focused on Azure cloud architecture for machine learning and data platform use cases
- Ph.D. in Bioinformatics and Computational Biology
- UNC Charlotte researcher
  - Infectious diseases (*E. coli* and *P. falciparum*)
  - Human genomics (Bantu and rare diseases)

parEBEN

StrainHub

Sparkitecture


BLUEGRANITE

# Solution Areas

*cloud scale analytics*

## Data Platform

Cloud-based, scalable solutions

## BI & Analytics

Self-service and enterprise

## AI & Machine Learning

Leading edge data science

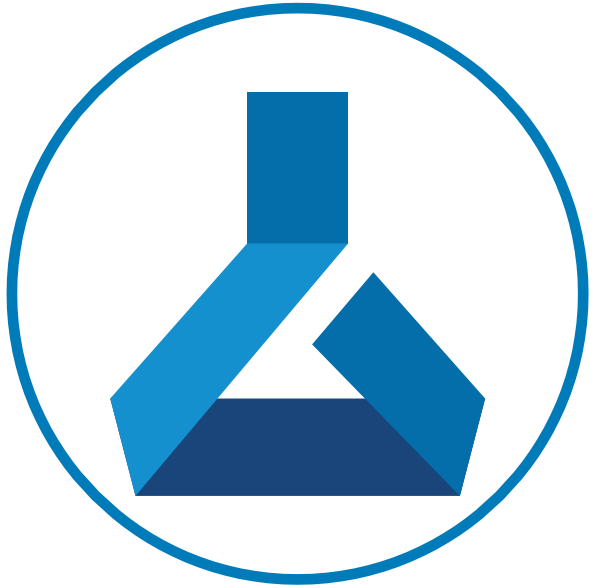
BLUEGRANITE

# Azure Services for Genomics-Based Analysis

## Databricks

Apache Spark platform for distributed computing

Includes: Databricks Runtime for Genomics + *Glow*



## Machine Learning Service

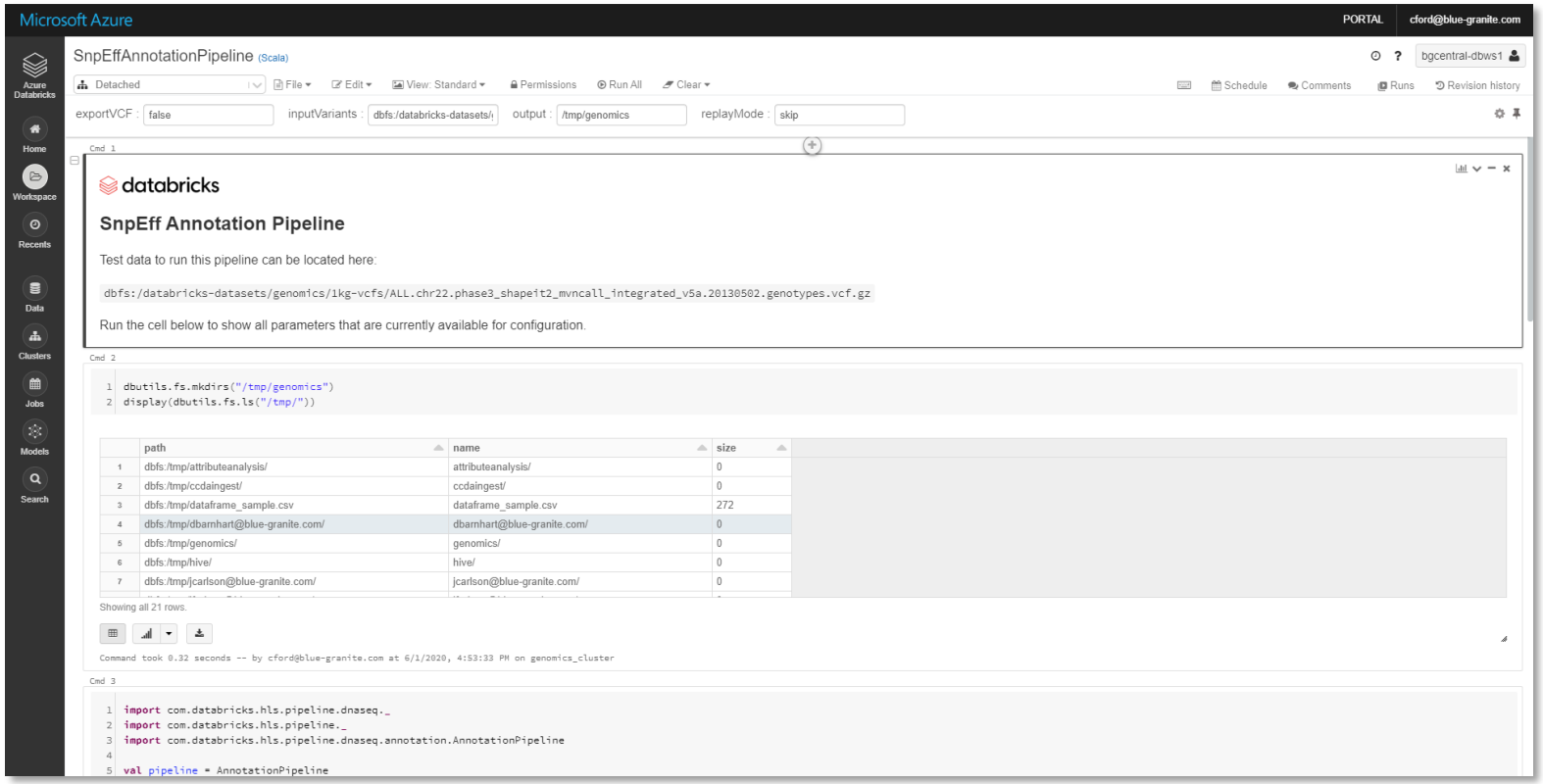
Familiar Jupyter or RStudio environment for interactive analyses and operationalization

## Genomics Service

Automated GATK-Compliant Processing Pipeline

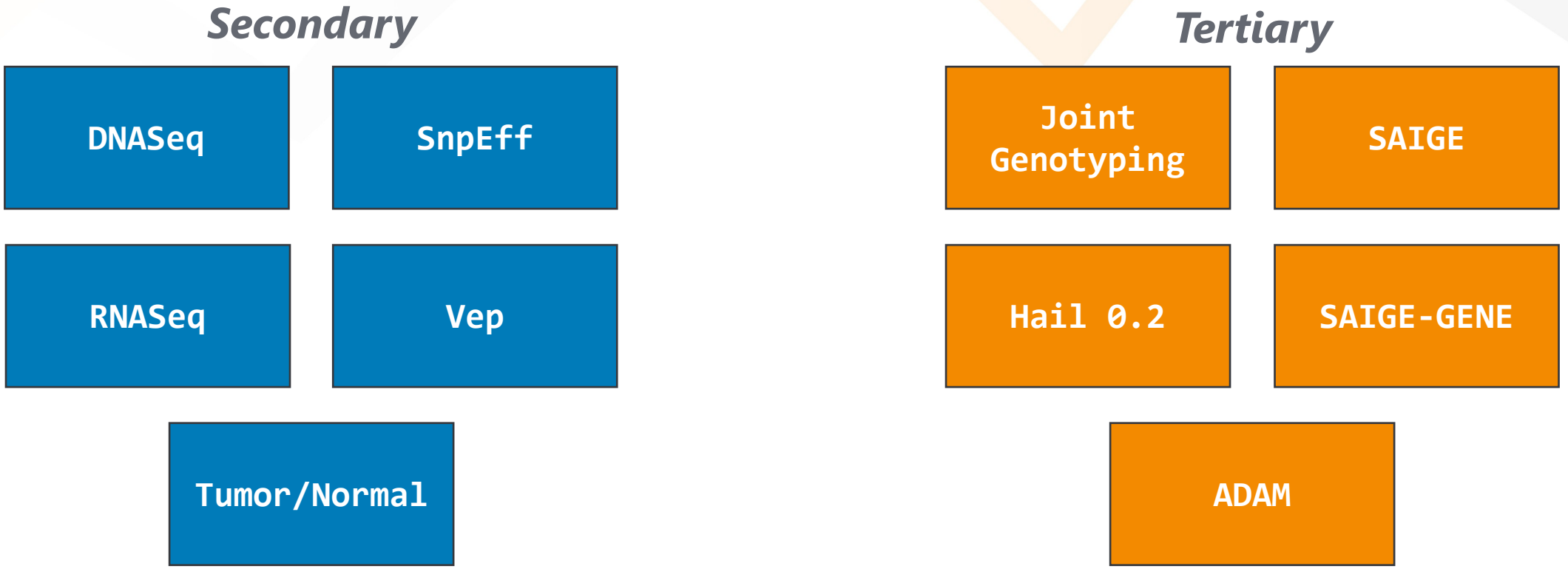
# Azure Databricks

Fast, easy, and collaborative Apache Spark™ based analytics service.




- Familiar notebook-style IDE
- Massively scalable
- Collaborative
- Python, R, Scala, and SQL

# Genomics Runtime - Analysis Pipelines




GLOW

# GLOW

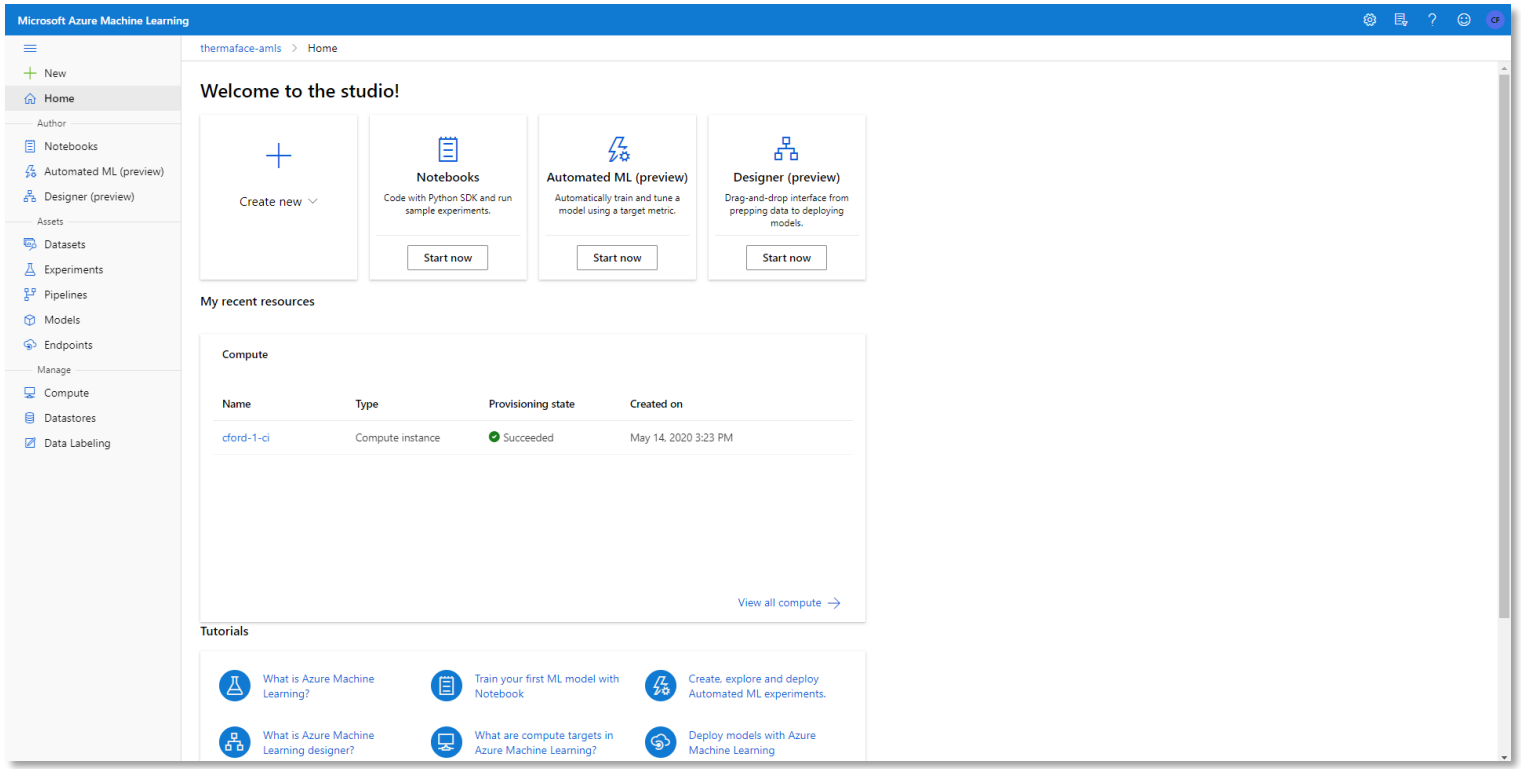
***Analysis Tools***

- Read and Write *VCF*, *Plink*, and *BGEN* Files
- Read Genome Annotations (*GFF3*)
- Perform Variant Quality Control
- Perform *liftOver* Genomic Conversions
- Perform Variant Normalization
- Split Multiallelic Variants
- Prepare Genomic Data for Machine Learning

- Parallelize Common Tools with a Transformer
- Utilize Python Statistics Libraries
- Perform GWAS Regression Tests

# Azure Machine Learning Service

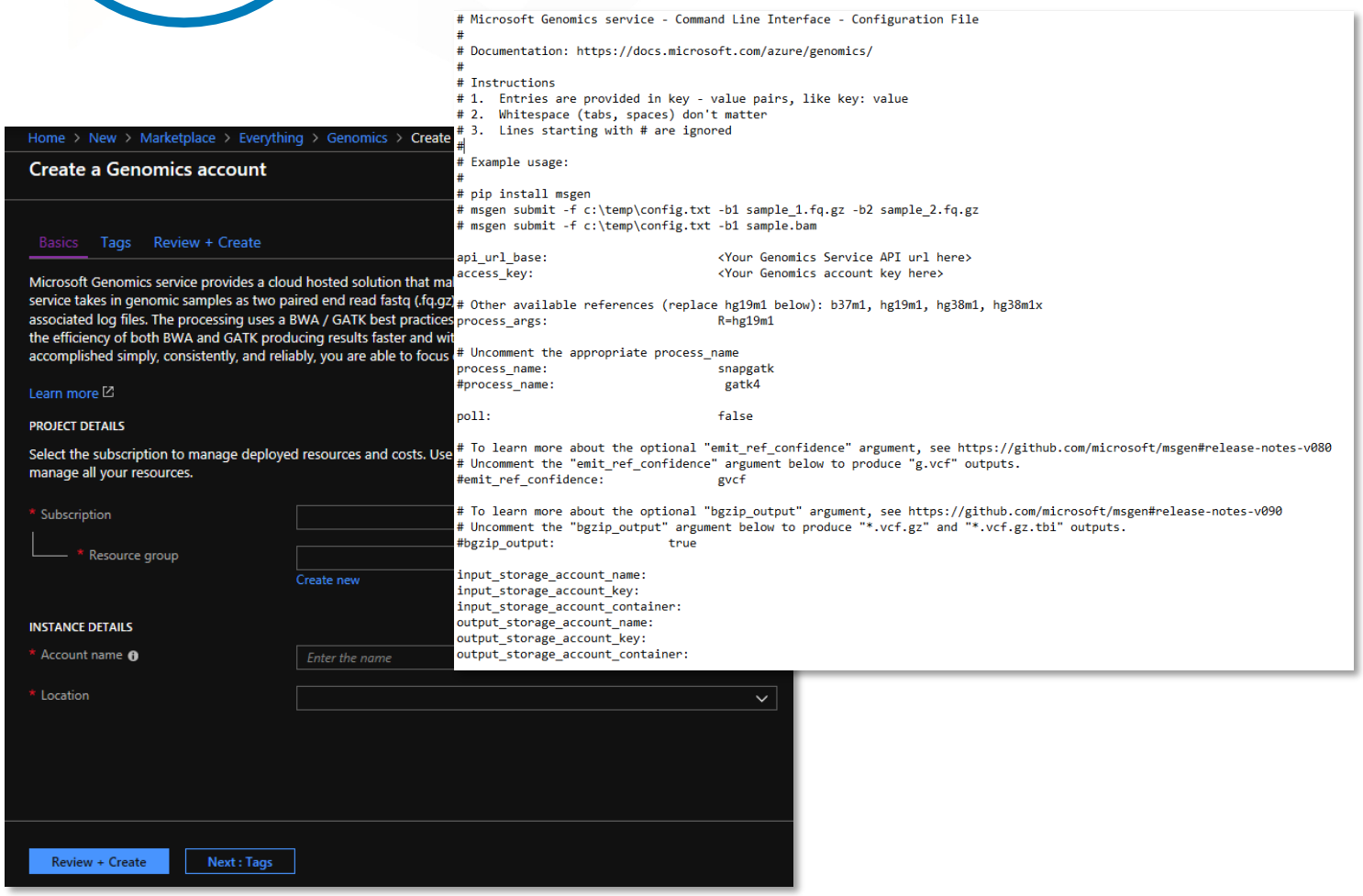
Enterprise-grade machine learning service to build and deploy models faster.




- Familiar JupyterLab, Jupyter, and RStudio IDEs
- Python and R SDKs
- Easy operationalization of code as APIs

# Azure Genomics Service

Power genome sequencing and research insights.

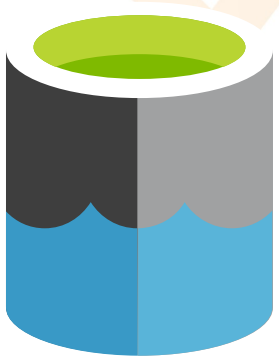



```
# Microsoft Genomics service - Command Line Interface - Configuration File
#
# Documentation: https://docs.microsoft.com/azure/genomics/
#
# Instructions
# 1.  Entries are provided in key - value pairs, like key: value
# 2.  Whitespace (tabs, spaces) don't matter
# 3.  Lines starting with # are ignored
#
# Example usage:
#
# pip install msgen
# msgen submit -f c:\temp\config.txt -b1 sample_1.fq.gz -b2 sample_2.fq.gz
# msgen submit -f c:\temp\config.txt -b1 sample.bam

api_url_base:                       <Your Genomics Service API url here>
access_key:                         <Your Genomics account key here>

# Other available references (replace hg19m1 below): b37m1, hg19m1, hg38m1, hg38m1x
process_args:                       R=hg19m1

# Uncomment the appropriate process_name
process_name:                       snapgatk
#process_name:                       gatk4

poll:                               false

# To learn more about the optional "emit_ref_confidence" argument, see https://github.com/microsoft/msgen#release-notes-v080
# Uncomment the "emit_ref_confidence" argument below to produce "g.vcf" outputs.
#emit_ref_confidence:                gvcf

# To learn more about the optional "bgzip_output" argument, see https://github.com/microsoft/msgen#release-notes-v090
# Uncomment the "bgzip_output" argument below to produce "*.vcf.gz" and "*.vcf.gz.tbi" outputs.
#bgzip_output:               true

input_storage_account_name:
input_storage_account_key:
input_storage_account_container:
output_storage_account_name:
output_storage_account_key:
output_storage_account_container:
```

- Python 2-based API
- Cloud implementation of Burrows-Wheeler Aligner (BWA) and the Genome Analysis Toolkit (GATK) for secondary analysis
- Uses FASTQ or BAM inputs

⚠ This service is a bit outdated and unmaintained by Microsoft at this point

# *But what about storing and organizing data in its many forms?*



***Azure Storage (Data Lake)***

Petabyte-size file storage with RBAC, auditing, and massive scalability

+

***Delta Lake***

Open-source storage layer that brings ACID transactions to big data workloads.

# Data Lake Zones

eCommerce Data

Devices & Sensors

Social Media

Web Logs

Corporate Data

Master Data

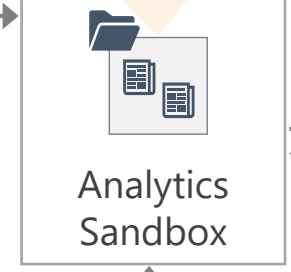
User Data

Landing Data Zone

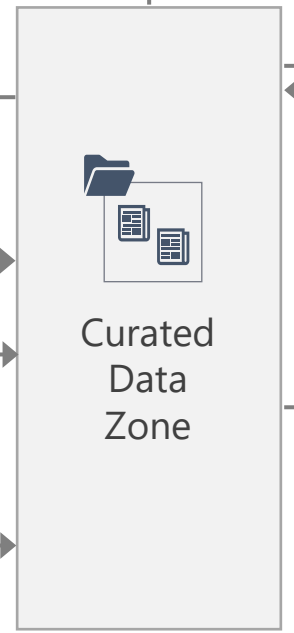
Archive Data

Data Warehouse

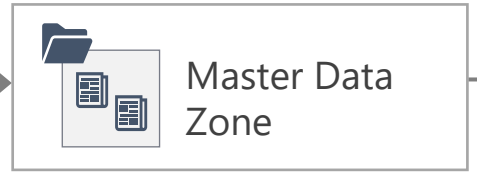
Raw Data Zone



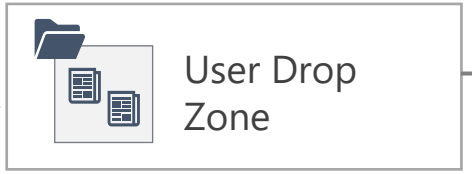
Analytics Sandbox



Curated Data Zone



Master Data Zone



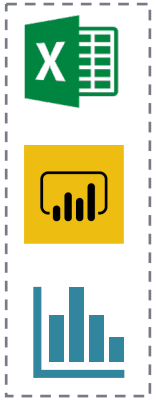
User Drop Zone

Advanced Analytics & Data Science

Machine Learning

R, Python, APIs

Prebuilt AI



Analytics

Data Exploration

Corporate Reporting

Self-Service BI

# Data Lake Zones (-omics)

DNA/
RNAseq

miRNA

Exomics

$CH_3$
Methylation

Phenotypes

EMR/Clinical
Data

User
Data

Landing Data
Zone

Raw Data
Zone

Master Data
Zone

User Drop
Zone

Archive Data

Analytics
Sandbox

Processed
Data
Zone

Data Warehouse

Advanced Analytics
& Data Science

Machine Learning

R, Python, APIs

Prebuilt AI

Analytics

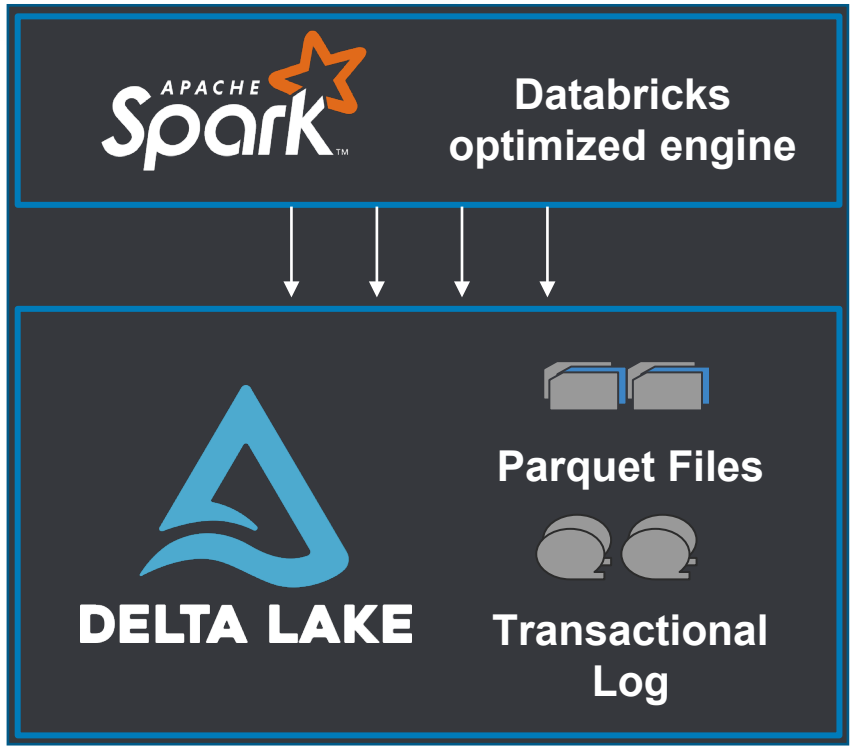
Data Exploration

Clinical Reporting

Self-Service BI

# Delta Lake Format

New Open Source Format Based on Parquet

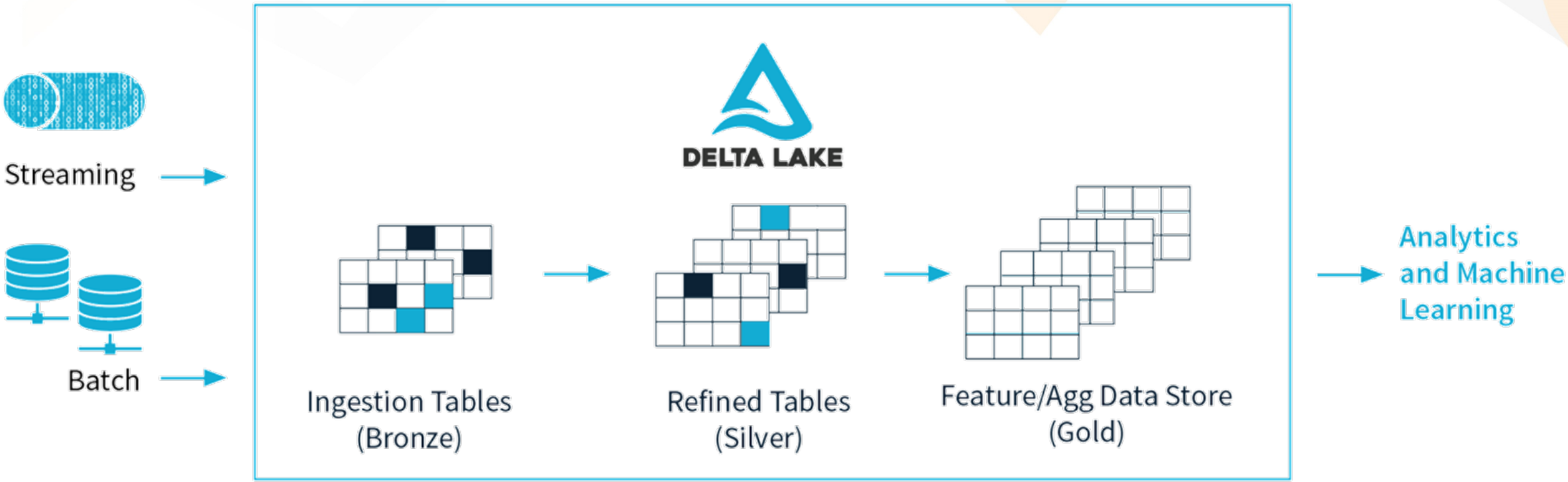



Highly Performant queries at scale

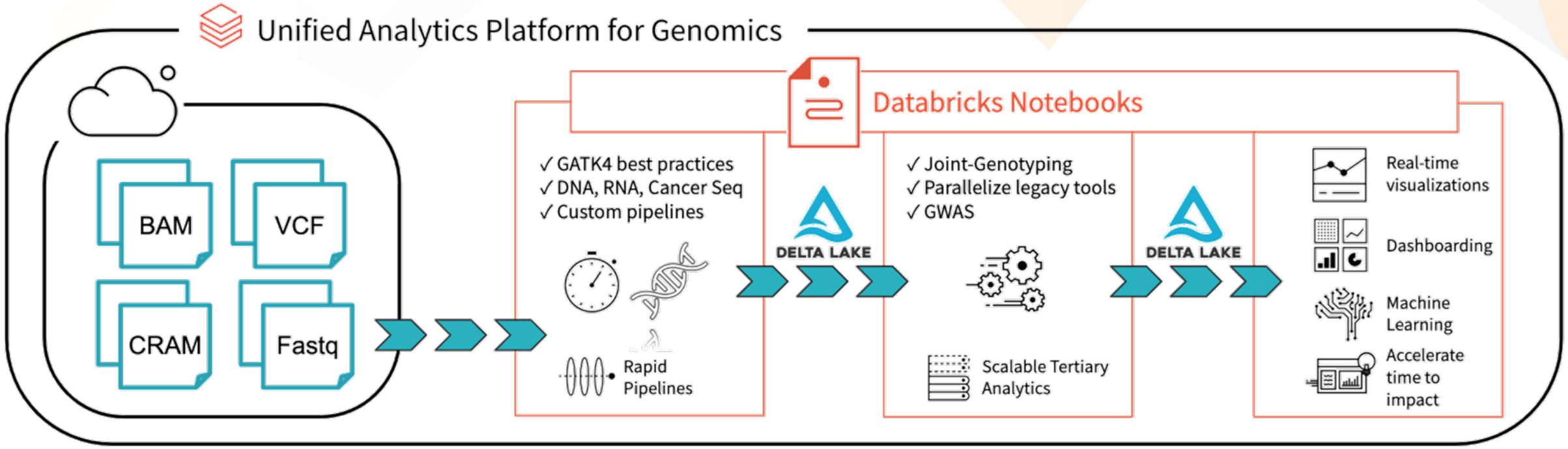
High Quality & Reliable Data always ready for analytics

**Key Features**

- ACID Transactions
- Schema Enforcement
- Unified Batch & Streaming
- Time Travel/Data Snapshots
- Indexing
- Compaction
- Data skipping
- Caching


Streaming
Batch
DELTA LAKE
Ingestion Tables
(Bronze)
Refined Tables
(Silver)
Feature/Agg Data Store
(Gold)
Analytics
and Machine
Learning

# Delta Lake Workflow and Architecture



# Delta Lake Workflow and Architecture