BlueGranite - Genomics

BlueGranite

DIGITAL TRANSFORMATION IN GENOMICS

GENOMICS DATA LAKE

CENTRALIZED DATA

Improve the accessibility of your genomics data by breaking down the silos and putting your data in a central location. Data lakes provide scalable and performant access for cloud-based analyses.

SCALE AND AUTOMATE

FASTER ANALYSES

Reduce processing times for secondary and tertiary analyses by harnessing the power of scalable computing services in the cloud. Automate processing pipelines to streamline your workflows.

IMPACTFUL REPORTING

SHOWCASE INSIGHTS

Showcase research insights and project performance by creating interactive dashboards and reports with Power BI. Facilitate improved access and knowledge sharing of your work.

Featured Solution Brief

BLUEGRANITE SCALES COLLABORATIVE RESEARCH WITH THE CREATION OF A GENOMICS DATA LAKE

When a $30M research endeavor plans to create over 400TB of multi-omics data, the cloud is the obvious option for scale and performance. A large research organization out of the Southeastern U.S. partnered with BlueGranite to provision a secure environment to house their genetic data. Built using Azure Data Lake and Azure Data Factory, we can collect data from constituent research groups and allows for the secure management and control over the data assets in the data lake. Plus, future enhancement will include scalable analyses using Azure Databricks to gain insights from this massive amount of human health information.

Read More

OVERVIEW:

- Collaborative Framework: Working with academic research groups and enterprise IT architecture, we created a solution for uploading and using data while retaining security.

- Solution Design and Security: The team worked within the constraints of the Azure Government Cloud to create a scalable genomics data lake while ensuring NIST 800-171 compliance for data security.
- Data Lake-Centric: For this solution, exome sequences along with phenotypes, proteomics, methylomics, and more needed to be logically organized for future cohort-based analyses. By using Azure Data Lake, the heterogeneous data was organized and cataloged at scale.

GENOMICS BLOG POSTS

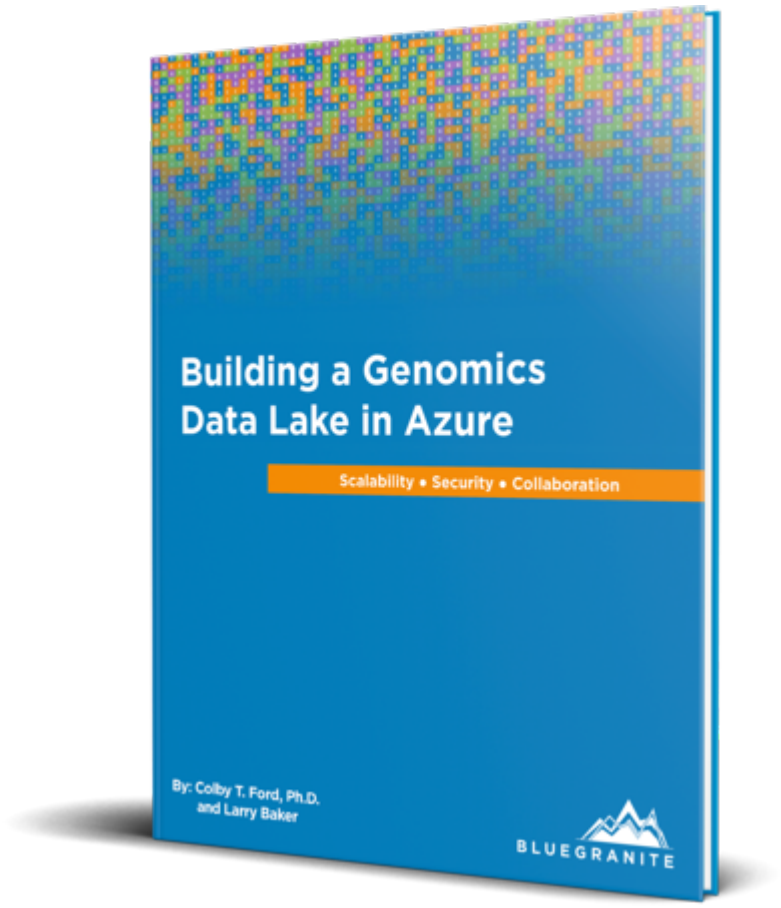
Explore the BlueGranite team's insightful blog posts on bioinformatics, genomics, and life science topics below:

BlueGranite eBook

Building a Genomics Data Lake in Azure

Scalability in genomics starts with a performant, secure, and collaborative space to store your data. In this eBook, we cover the ideas around building a data lake for your genomics data, including organization, security, and automation of analyses.

Download the eBook



BlueGranite Solution

Azure Data Factory Connector for Illumina®

BaseSpace®

Easily copy your data from your BaseSpace account over to your Genomics Data Lake in Azure. This automated approach for retrieving your project samples, analysis outputs, and other datasets unlocks the ability to take advantage of the Azure cloud for secondary and tertiary analyses, machine learning, and more.

SOLUTION OVERVIEW:

- Deployment of an Azure Data Factory pipeline to securely copy data from your Illumina® BaseSpace® account to Azure Data Lake.
- Automation of this pipeline will retrieve new data as it becomes available, organizing the data by Project, Run, Dataset, and data type.

Projects	
Samples	.bcl, .fastq
Analysis Results	.bam, .vcf
Other Datasets	.csv, logs, etc.

Azure Services for Genomics

Databricks Runtime for Genomics + GLOW

Massively scalable, fast, and collaborative Apache Spark™-based analytics service.

- Familiar notebook-style IDE with Python, R, Scala, and SQL
- Easily read VCF and BGEN files into Spark DataFrames
- Perform secondary and tertiary analyses at scale
- GloWGR for large-scale regression analyses

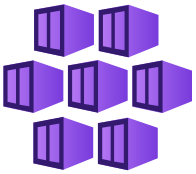
Learn More

Azure Machine Learning

Scalable workspaces for machine learning and bioinformatics experiments.

- Familiar JupyterLab, Jupyter, and RStudio IDEs
- Python and R SDKs
- Easy operationalization of code as APIs
- Easily use Bioconductor or other packages

Learn More

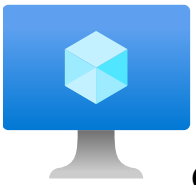


Azure Kubernetes Service

Serve scalable compute resources in Docker containers of virtually any application.

- Scale up or scale out for faster processing
- Perfect for porting over HPC workloads to the cloud
- Tools for continuous integration and continuous deployment (CI/CD) workflows

Learn More



Genomics Data Science Virtual Machine

Prebuilt virtual machine image with pre-installed software for bioinformatics and ML.

- Python, R, Bioconductor, Spark, JupyterHub, and RStudio
- H2O, XGBoost, TensorFlow, etc.
- Microsoft Genomics Jupyter Notebooks
- GATK 4.1.8.1

Learn More

Microsoft Genomics Service

Automated GATK-compliant pipeline for sequence alignment and annotation.

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

Learn More

BlueGranite Solution

Power BI for Bioinformatics

Create interactive dashboards and reports of your genomics data with Power BI. By using our expertise coupled with some Power Query magic, we can read and visualize all sorts of files that are common in bioinformatics.

This enables users to take advantage of information in files such as .FASTQ, .BAM, .VCF, and .GFF. Also, you can now import data from virtually any site such as the Protein Data Bank, NCBI, PlasmoDB, and more.

Check out our Demo Video:

FEATURED CONTACT

Colby Ford

Colby is Blue Granite's Principal of Life Sciences. He helps clients in this space envision solutions that advance data management and production, speed insight delivery, to improve business outcomes. Colby's specialties include AI in genomics, phylogenetics, protein structure modeling, and the design of scalable bioinformatics pipelines in Azure.

VIEW LINKEDIN PROFILE

SCHEDULE A CALL

Feel free to schedule a call to discuss any of the content above
or any questions you may have. I am happy to help!