Enabling Faster and More Cost-Effective Data Management Using flexFS™ with the Allotrope Data Model

Sep 15, 2021
Agenda

• Digitization in Drug Substance Manufacturing
  – Prescriptive analytics
  – Workflows to optimize testing

• Overview of flexFS

• Use Cases with the Allotrope Data Format
  – Read/Write performance with LC-MS data
  – QC workflow at a single site
  – QC workflow at multiple sites with a Single Point of Truth (SPOT) data server

• Data storage and querying challenges

• Conclusions
Goal is Prescriptive Analytics

Descriptive Analytics
“What happened?”
Provides insights into the past. Example: Control charts of cell growth

Diagnostics Analytics
“What did it happen?”
Digging deeper. Example: Chromatogram review, Batch-scale comparison model

Predictive Analytics
“What will happen?”
Forecasting and understanding the future. Example: real-time MVDA process monitoring

Prescriptive Analytics
“How can we make it happen?”
Optimization and simulation for possible outcomes. Example: Smart automation for optimized setpoints and controls

Shum outlines the digital environment

We will explore elastic file store
Previously showed elastic compute

Data Sources
- Data Source 1
- Data Source 2

Calculation Engines
- R; Python; Others

Virtual Machines/Applications
- Calculation Modules
- Off the Shelf Software

Data Connections
- API’s

Analytics Environment

Easy Access User Interface
Shum suggests workflows to optimize testing

Require scalable compute and easy access to data

1. Data Collection
   - Failure Cycles
   - Good Cycles

2. Generate Overlays

3. Determine Optimal Threshold

4. Evaluate Accuracy

5. Compare Methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy (%)</th>
<th>Prediction (%)</th>
<th>Failures Detected (%)</th>
<th>Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>95</td>
<td>35</td>
<td>90</td>
<td>Excellent</td>
</tr>
<tr>
<td>B</td>
<td>91</td>
<td>25</td>
<td>80</td>
<td>Good</td>
</tr>
<tr>
<td>C</td>
<td>94</td>
<td>30</td>
<td>60</td>
<td>Good</td>
</tr>
<tr>
<td>D</td>
<td>84</td>
<td>30</td>
<td>50</td>
<td>Poor</td>
</tr>
</tbody>
</table>
With optimal testing – acceptance testing can be established

1. “min” notation means to take the lesser of the two values. This reduces over-estimation.

2. Assumes a normal distribution (not true).

Challenges to make scalable compute cost effective

Data Analysis & Querying Challenges

Challenges to meet at minimal cost:

- How do you scale for machine learning?
  - Goal is in the range of millions to billions of calculations
  - Algorithms needed to be scalable
- Storage of different files – redundancy of data in different file formats (solved with ADF, thank you)
- Analysis results needed to be readily available and easily queryable
  - For multiple users
  - From multiple sites
- Data governance –
  - ACID, extended ACLs (beyond POSIX)
  - Encrypted

Interactive querying
- Metadata
- Annotations
- UV-LC
- LC-MS
- Results data
- Transactional (ACID)

Analytics built in – R and Py APIs
REVEAL™ Architecture
in the cloud: no code R API, combined with array native indexing, elastic compute and elastic file storage

REVEAL Suite of Apps

SciDB
Primary Cluster
Always Online
Primary Volumes

Burst Mode™
Elastic Task Service
Spot Workers
Minimal Storage
CPUs or GPUs

flexFS™
Network Filesystem
Large Datasets

Works with cloud management services
Architecture

Filesystem Host
- mount.flexfs executable
- FUSE (userspace filesystem)

Network Transport
- Transmissions encrypted with TLS
- Optional end-to-end encryption (AES-256)

Admin Service
- Manage users and volumes
- SaaS and self-hosted options available
- Highly available

Metadata Service
- Low latency
- SaaS and self-hosted options available
- Highly available

File Block Storage
- Pluggable architecture
- AWS S3, GCS, Azure Blob, MinIO, OpenIO, Wasabi, etc.
- Encryption at rest (AES-256)
SciDB+ flexFS + SciDB Bridge

High-scalability shared arrays accessible to multiple SciDB clusters and burst workers

SciDB Cluster 1
- Bridge Storage (flexFS)
  - Bridge Arrays
    - Metadata
    - Index
    - Chunks (Arrow)

SciDB Cluster 2
- Bridge Storage (flexFS)
  - Bridge Arrays
    - Metadata
    - Index
    - Chunks (Arrow)

Reference Storage (S3)
- Bridge Arrays
- Files

Burst Workers
Despite there now being 5 concurrent clients, each client gets roughly the same throughput with AWS S3 CP and flexFS as it would if it were the only client. However, EBS (via SSHFS) and EFS both show roughly 5x decreases in per-client throughput. This decrease in throughput is linearly proportional to the (actively loaded) cluster size.
Use Case 1 – Read/Write Performance (LC-MS Data)
Operations performed on a ~6 GB LC-MS HDF5 file

- Plot shows the multiples by which flexFS outperforms EFS
- Increase in relative performance of flexFS with cluster size
- c5n.18x large EC2 instance
Use Case 2 – Single-Site QC Workflow
a la Dr. Shum’s paper

- **Data ingestion of ADF files**
- **Encryption in transit**
- **POSIX compliant file system**
- **S3 backed**
- **Unlimited data storage**
- **Encryption at rest**
- **Cost-effective storage**

**User 1**
- flexFS backed

**User 2**
- flexFS backed

**User 3**
- flexFS backed

**Instrument 1**
**Instrument 2**
**Instrument 3**

**ADF Files**
**flexFS Data Server**
**ADF I/O API’s**

- **Highly performant data querying**
- **Encryption in transit**
Use Case 3 – Multi-Site QC Workflow

DB + Analytics Layer/ REVEAL Apps

- Single Point Of Truth (SPOT)
- flexFS Data Server

- In-Database Computations
- Aggregate Data

- Support multiple sites
- High performance for concurrent operations
- Easy data availability for reuse

Site 1
- LC-MS-1
- LC-UV-1
- LC-UV-2
- ADF Files

Site 2
- LC-MS-1
- LC-UV-1
- LC-UV-2
- ADF Files

User 1
flexFS backed

User 2
flexFS backed

User 3
flexFS backed

User 1
flexFS backed

User 2
flexFS backed

User 3
flexFS backed
Storage Challenges with Data Growth

Data Generation, Storage, Accessibility

- Raw; Vendor Format – e.g., 2 GB/file
- LC-MS Files raw/vendor
- Study can have associated pathology images – e.g., 10 GB/file

- Unpacked; Versioned; Processed \( \rightarrow \sim 3-5 \times \) expansion – e.g., 10 GB/file
- LC-MS Files unpacked
- Site 1-Lab 1
- Site 1-Lab 2
- Site 2-Lab 1

10's of TB's of data generated/year
- Storage required over time – raw, processed, analyzed
- High accessibility required to data
- Statistical algorithms helpful on central data server (SPOT)
- ADF format vital for federation

Current Landscape of Storage Solutions

Comparison of Storage Costs per Month for 50 TB

- Estimates for storage of 50 TB of file data, actively accessed
- Comparisons made using multiple-availability-zone pricing
- Same region data storage & access (no transfer charges)
- S3 API costs, while not necessarily trivial were not considered
- On-demand pricing discounted by 30% is used for all server cost calculations
- File and metadata servers are all normalized to m5.2xlarge instances for consistency
flexFS Installation and Usage

Repeat on as many servers as desired.

# 1) Install FUSE userspace tools if missing.
$ sudo yum install -y fuse

# 2) Download and install the flexFS mount client
$ curl http://admin.flexfs.io/<path_to_dir>/mount.flexfs -o mount.flexfs
$ chmod +x mount.flexfs
$ sudo mv mount.flexfs /sbin

# 3) Initialize the flexFS volume
$ sudo mount.flexfs init --admin-addr <yourdomain>.admin.flexfs.io:443

# 4) Mount the flexFS volume
# Replace <mountpoint> with relevant folder (e.g. /flexfs/<mountname>).
$ sudo mount.flexfs start <mountname> <mountpoint>

- Show mounted flexFS volume on drive
- Example operations using h5dump, h5ls, h5copy on HDF5 file
  - View root groups and structure
  - View specific group information
  - View specific dataset
  - Create new file from subset
- File transfers between flexFS backed server and client machines
- m5.8x large EC2 instances
Conclusion

• Based on work by Dr. Shum and others, it’s clear that digitization will require
  – Elastic data management and
  – Elastic analysis

• There are many strategies for achieving elasticity

• We show the
  – Comparison of the scaling capability of
    • a posix compliant file store, flexFS, that works with block storage found in the cloud
    • Vs EFS found in the AWS cloud
    • The comparison in scaling showed the advantages of a solution with parallel read and write capabilities
  – The potential for savings using a file store, like flexFS for managing large file storage.
Thank You!