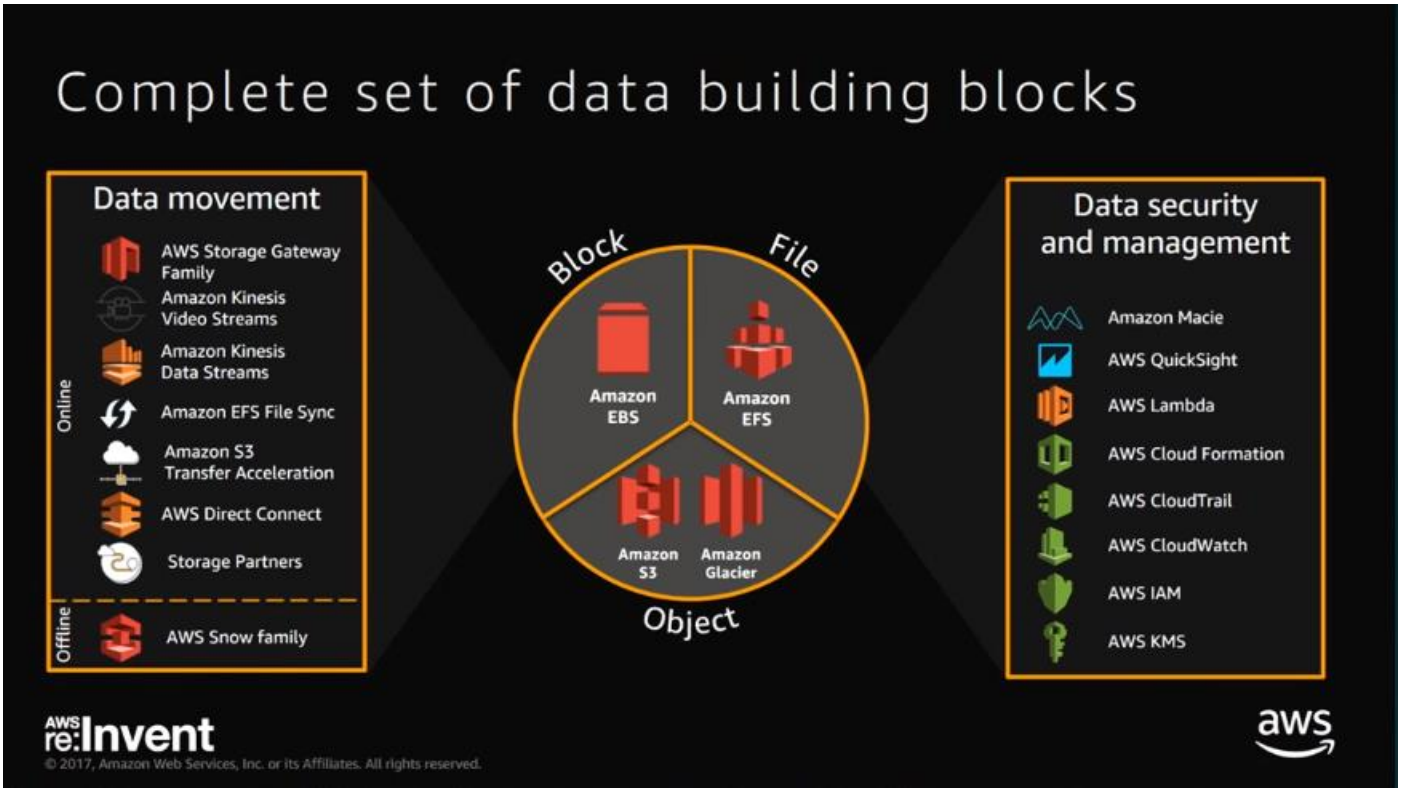
This is your chance to learn directly from top CTOs and Cloud Architects from some of the most innovative AWS customers. In this lightning round session, we'll have an action-packed hour, jumping straight to the architecture and technical detail for some of the most innovative data storage solutions of 2017. Hear how Insitu collects and analyzes data from drone flights in the field with AWS Snowball Edge. See how iRobot collects and analyzes IoT data from their robotic vacuums, mops, and pool cleaners. Learn how Viber maintains a petabyte-scale data lake on Amazon S3. Understand how Alert Logic scales their massive SaaS cloud security solution on Amazon S3 & Amazon Glacier.



What's in store...

7- to 10-minute highly-technical presentations

- **Alert Logic** – ingestion, storage & analytics
- **iRobot** – analyzing IoT data
- **Celgene** – application migration
- **Viber** – building a data lake
- **Insitu** – data at the edge

Alert Logic

Scalable Ingestion, Storage & Analytics

Paul Fisher

Technical Fellow

About Alert Logic

Alert Logic provides fully managed security monitoring and protection for cloud-deployed applications and workloads

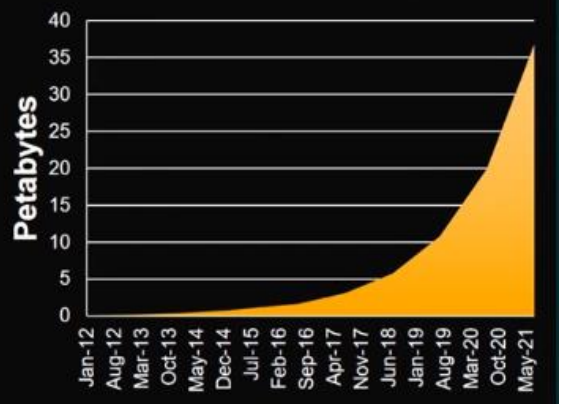
- Intrusion detection
- Web attack detection and blocking
- Vulnerability scanning
- Asset analysis and configuration assessment

ALERT LOGIC®

The problem we faced

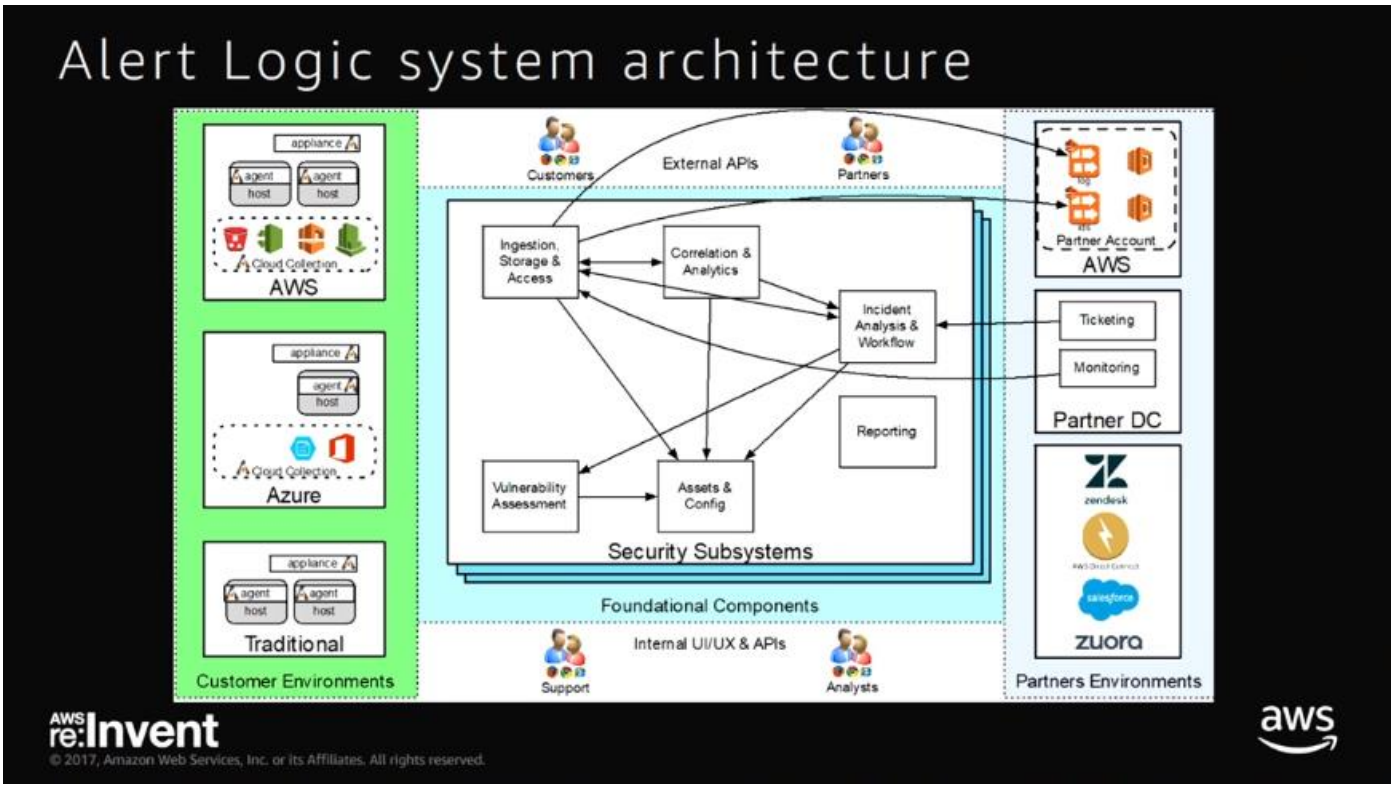
Doing this produced a large data-processing problem

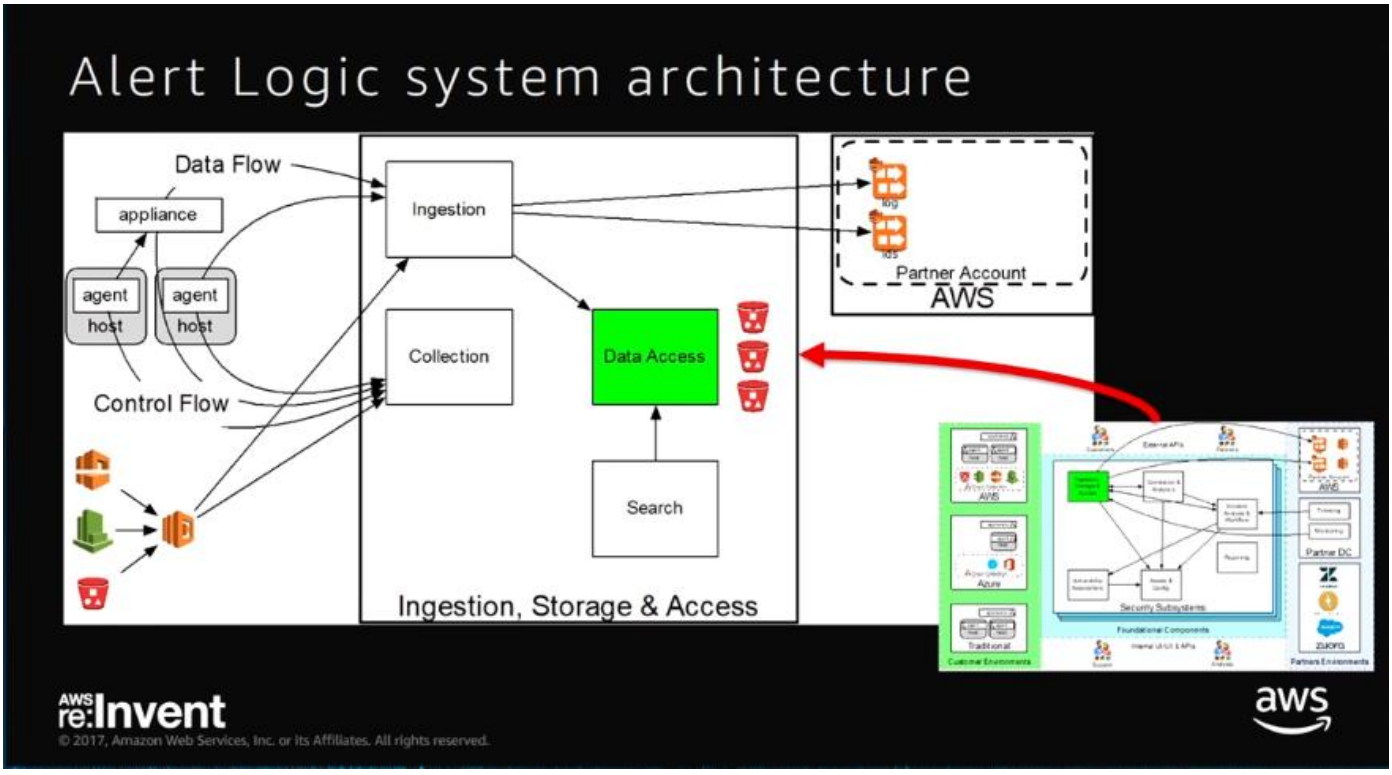
- 4,100+ customers
- 1.2M messages/second
- 2 petabytes per month
- 3 months to 7 years retention
- Adding 110 percent data volume/year



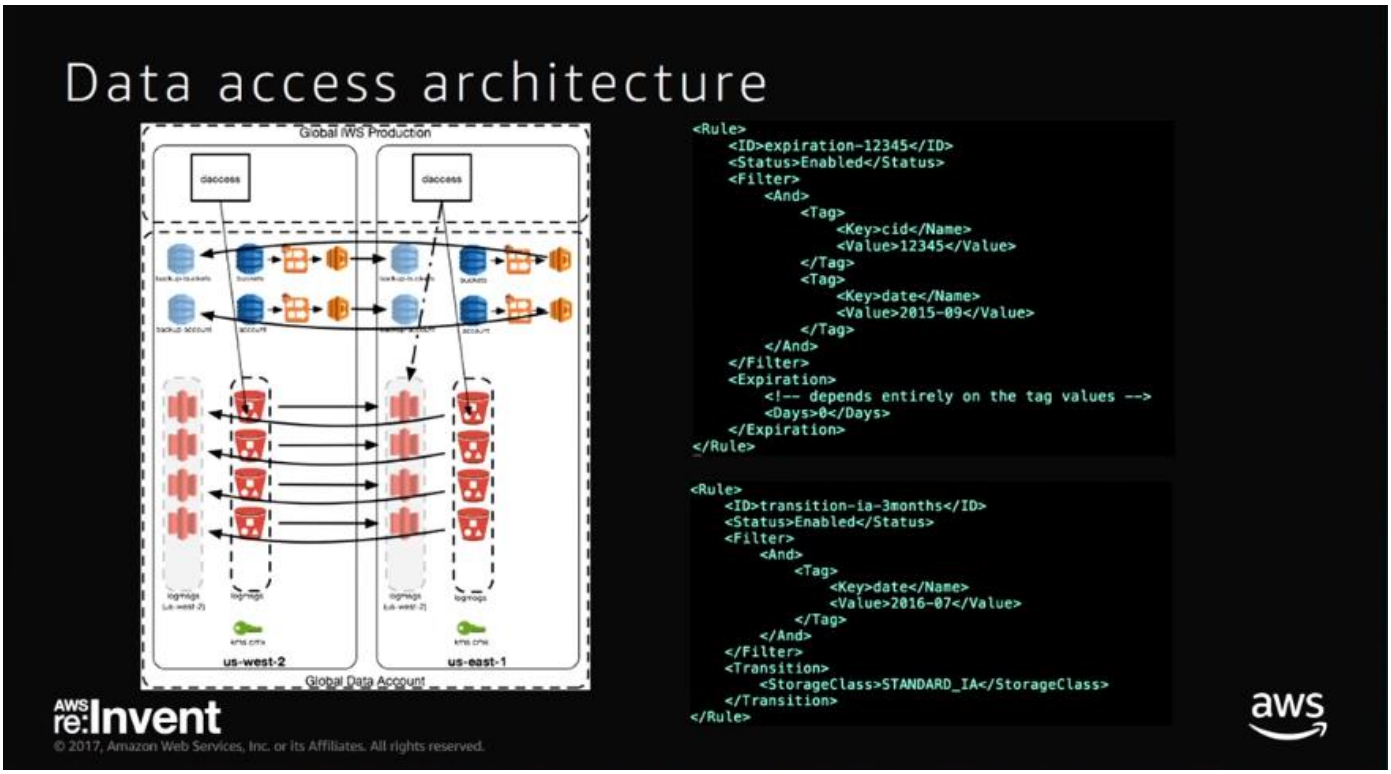
Key architecture challenges

- Achieving performance, durability, and availability, all at once, is hard
- Providing multi-region availability with Recovery Time Objective (RTO) of 15 minutes
- Designing for scale up 150x to 10PB per day for 100k customers
- Doing this all for less than $0.02 per GB, with a team of 20 engineers





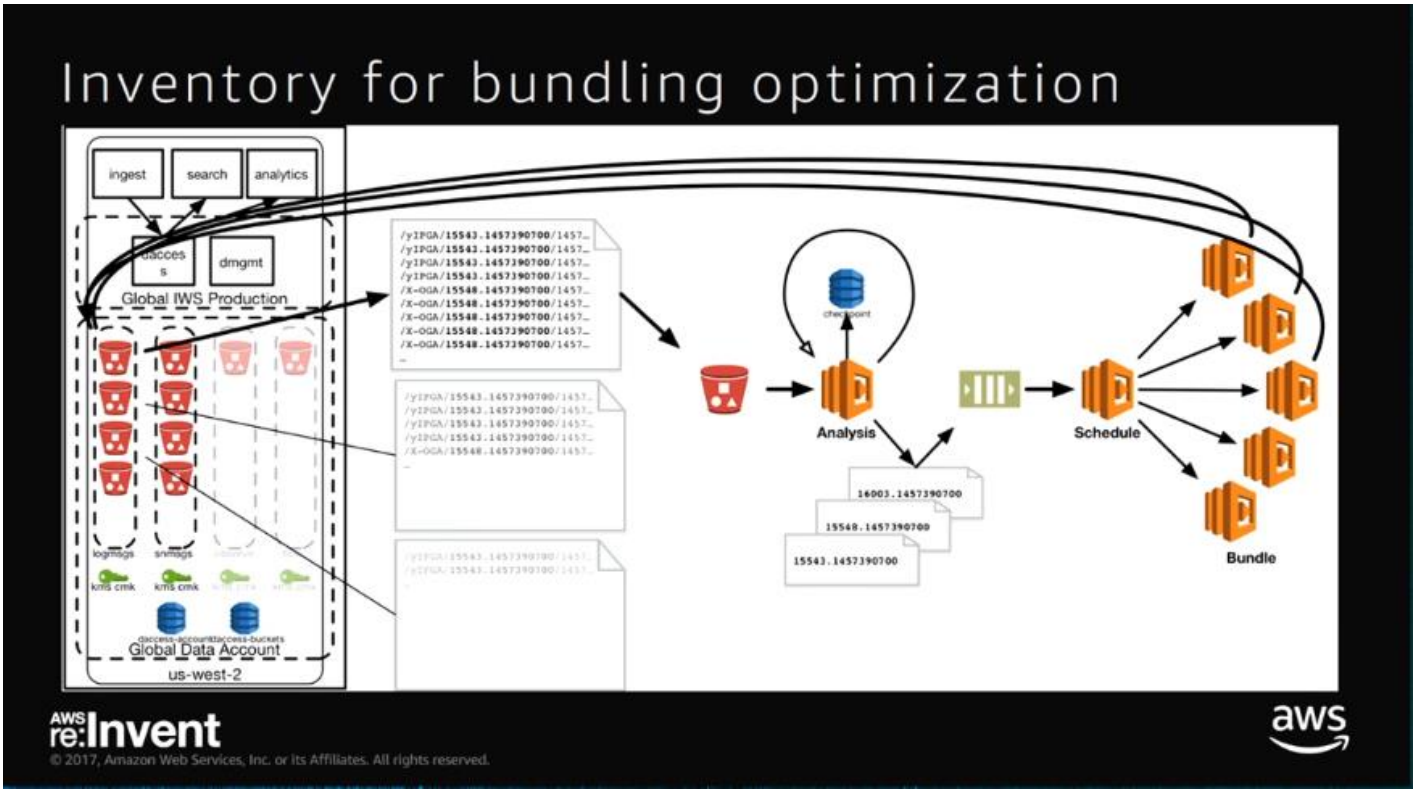
All the data management is built in a way that it is multi-region, high availability and is being done as part of its implementation described below



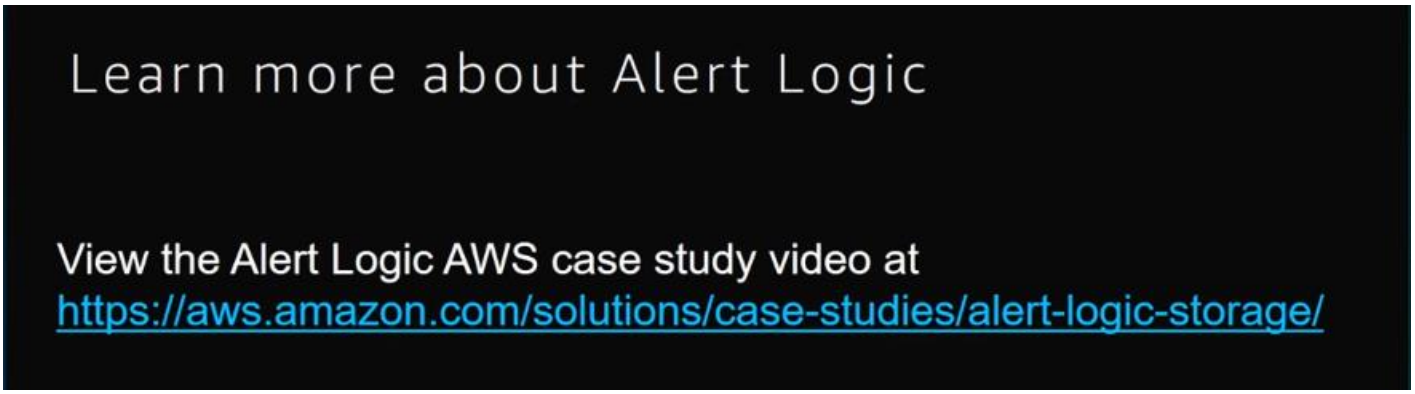
The left side of the diagram shows 2 AWS regions each with a set of S3 buckets for each data type that is being stored. The buckets are then arranged so that we put 950 customers per bucket because we need to do retention periods on a per customer basis. So, we write 950 life cycle rules for each customer per month so that we can expire the data after a month, you also write storage class tiering rules that allows us to be able to tier down the data into infrequent access, in the backup region, we then write that data to infrequent access directly and then tier it down to Glacier.

Those Rules all pivot around 2 Tags, the top Rule on the right is a specific customer's expiration rule and the bottom segment is the storage tiering rule for the primary region. The regions have customers as primaries and the data bi-directionally replicate. There is also a set of KMS keys, one per data type that is generated. We then use that to generate data keys that are used to encrypt all the data. the primary region uses the backup region's KMS keys to generate a data key. All those data keys are written so that we have one key per customer per month (to reduce the blast radius of an encrypted key to just one customer for one month), those keys are written in the primary region. We then use CRR to replicate all of the data files and the key files over to the backup region.

At the same time, there are DynamoDB tables that keep track of the buckets that are being created for a given data type and the customers signing in into those buckets. We then have those data also replicated in equivalent DynamoDB tables in alternate regions. This means that in cases of downtime, we can just redirect traffic to the backup region and start servicing requests as if nothing happened. This is our disaster recovery strategy and our active-active architecture.



We have a need to bring data in rapidly, get it put into the system, and then have it processed in real time. We have lots of single small files that we can then bundle up based on an inventory report that was analyzed by a lambda from a notification.



We are going to see a part of iRobot's IoT workload for big data analytics flow of their Roomba data

The problem...

- We needed a facility to support ad-hoc queries, through an API to report historical states for AWS IoT Thing Shadow attributes.
- Our initial design pattern was based on Amazon Elastic Map Reduce, Spark Streaming, Amazon DynamoDB, Amazon API Gateway, and AWS Lambda
- DynamoDB became costly for this use case, and we needed to find an alternative solution to meet cost and performance expectations.

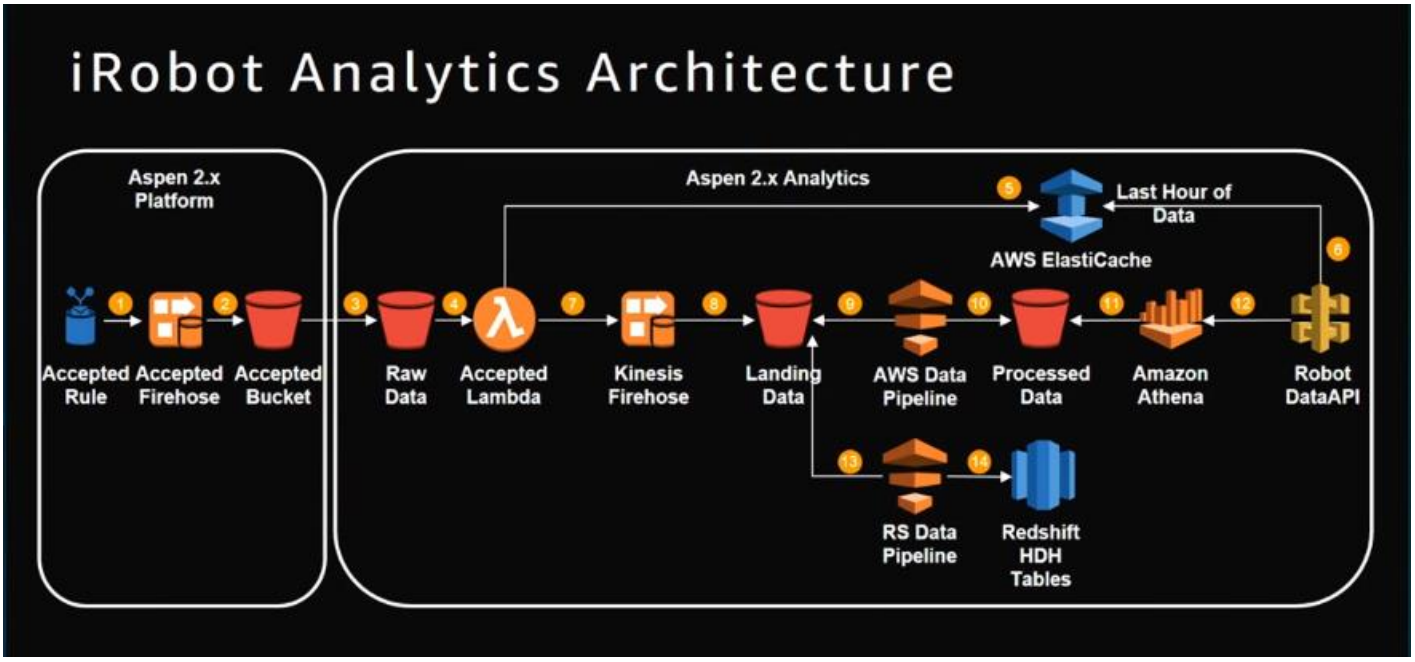
Key Architecture Considerations

Focused on AWS Managed Services

Balance Performance & Cost

Minimize DevOps impact

DynamoDB was the single largest cost component of the old system



We are using Rules created in ***AWS IOT*** to trigger off actions in places like ***Kinesis***. This production workload is designed to support 2 major use cases, the first is to support our customer support staff for customer issue resolutions by having the historical state of their device available when needed in real time through ***AWS ElastiCache*** path to the Robot API.

Linux migration

- Migrate COTS-HPC applications to AWS
- On-premise clusters no longer sized for data intensive projects
- Desire to move to elastic computational environment
- Data protections in place

2014
Amazon S3
Amazon Aurora
Amazon EC2

2015
Amazon RDS
AWS WAF
Amazon SQS
AWS Direct Connect
AWS CloudFront

2016
Amazon EFS
Amazon EMR
Amazon WorkSpaces

2017
AWS Storage Gateway
AWS Lambda

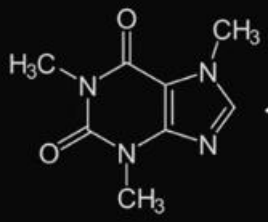
2018+
AWS
IoT

First Foray → First Migrations → Cloud First

Key architecture challenges

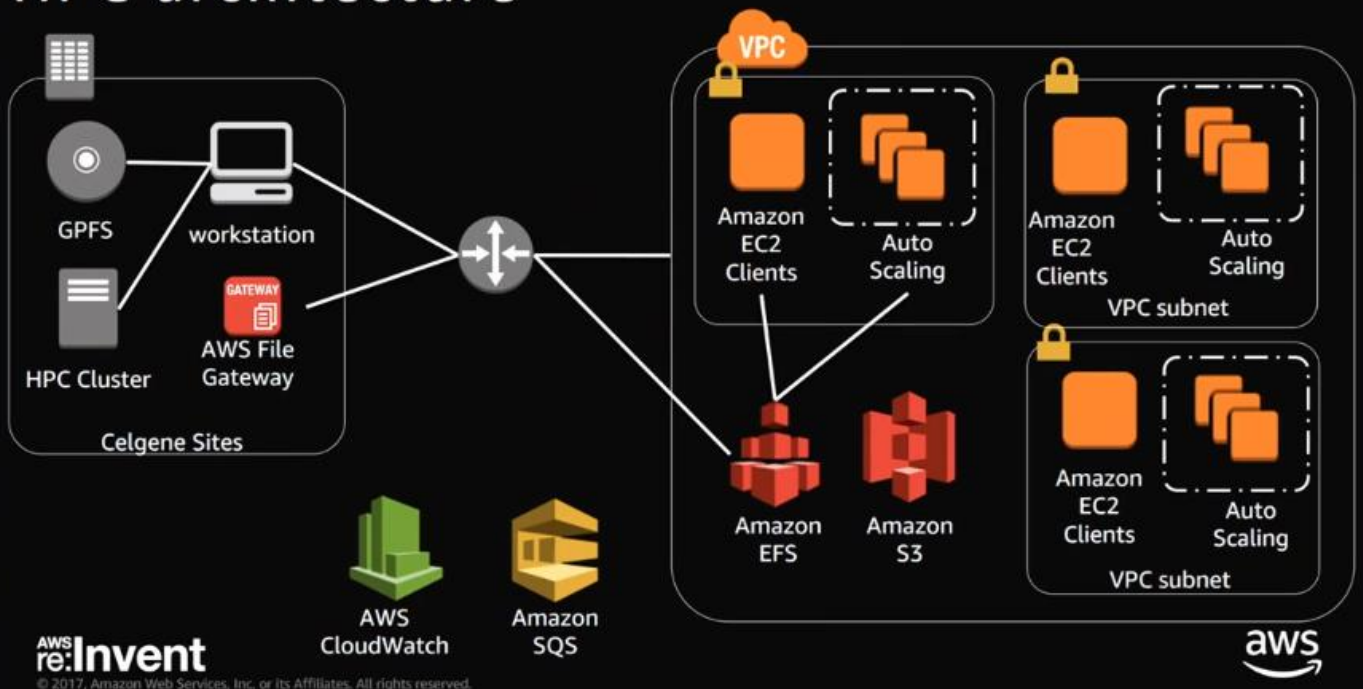
- Application not made for cloud
- No app/code changes
- Need to refactor for best practices

GENOMICS HPC



COMPUTATIONAL CHEMISTRY

HPC architecture



Hear more from Celgene...

STG310 – Files in AWS: Overcoming Storage Challenges for Common File Use Cases

Rakuten Viber

Viber—Data Lake Challenges

Amir Ish-Shalom

Chief Architect

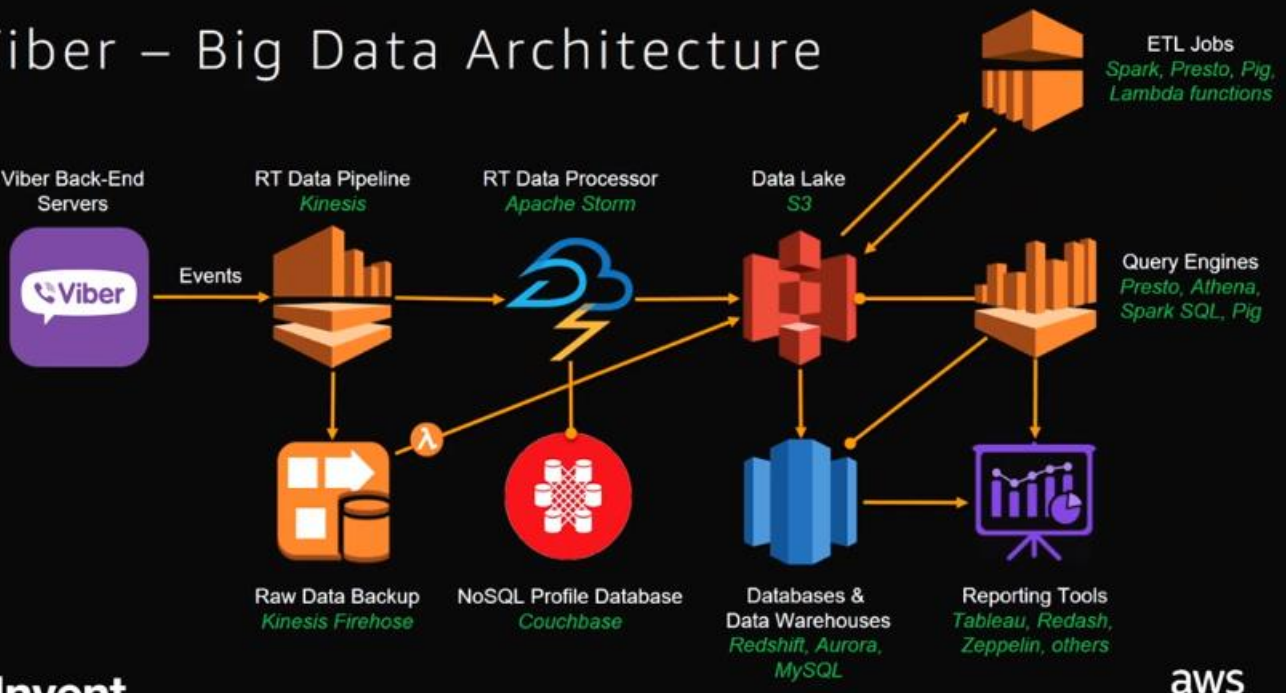
Rakuten Viber

- Messaging (incl. group)
- Secure end to end encryption
- Rich media & chat extensions
- Full multiple device support
- HD video & voice calls
- Viber out & Viber in
- Public chats & accounts
- Chatbots

Big Data @ Viber

- Close to 1 billion users worldwide
- Globally used in 230 countries
- 10-15 billion events daily (2 TB)
- 300,000 events per second (peak hours)
- 5 PB of data stored on S3/Glacier
- NoSQL DB (Couchbase) performing 2 million TPS on 20TB of data with 35 billion keys

Viber – Big Data Architecture



Use Case: S3 Performance

Challenge:

- Over 300 different event types with large throughput variance
- Storm data processor created many small files, especially for lower throughput events
- Events are stored in Hive partitioned folders (Y/M/D/H) which are not optimal for S3
- Running a query over these events using Presto could generate up to 15K tps on a single S3 bucket resulting in 5xx errors and throttling the whole bucket for other processes

Use Case: S3 Performance

Solution:

- Concatenate small files into large files, optimally 100MB+
- Convert files into columnar file format such as Parquet or ORC

Use Case: S3 Performance

Future Solution:

- Concatenate & convert files in a single process (Glue?)
- Use better partitioned hive directory format (H/D/M/Y instead of Y/M/D/H)
- Use even larger files for high throughput events

Learn more about data lakes

STG312 – Best Practices for Building a Data Lake in Amazon S3 & Amazon Glacier

STG313 – Big Data Breakthroughs: Query & Process Data In-place with Amazon S3 & Amazon Glacier

Yes Problemo!

- We are in the middle of nowhere – no Internet, think Australian Outback, an open-pit mine, a collection of gas well heads, linear infrastructure like pipelines
- We want to build an accurate (under 10 cm/pixel) 3D (or what we also call 2.5D) model - Why? Because our clients want our spatio-temporal algorithms to find what, if anything is broken
- We get 30+ TB of data for that location, we may have many many such locations operational globally

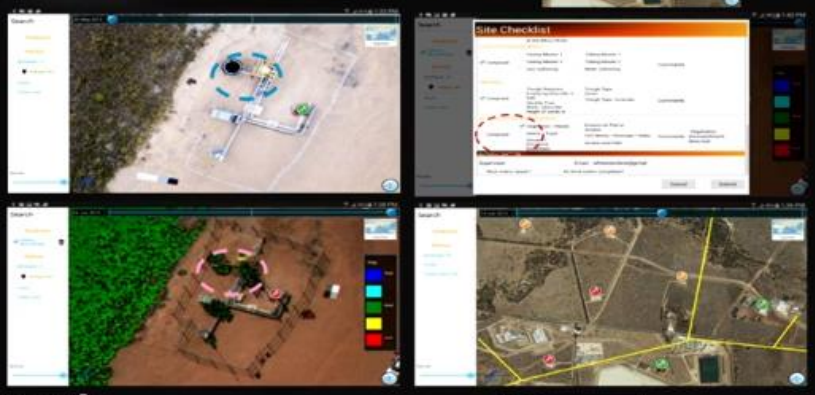
Problem

- We want to verify that the data we collected on that day was decent on the ground at the "edge"
- We want that data to reach our cloud for further and continuous future processing

Is my data good? How to I move it?

1

2



- Middle of nowhere
- Risk for people
- Less area covered
- Lot of data (TB)
- Time to result in days
- $$$$$

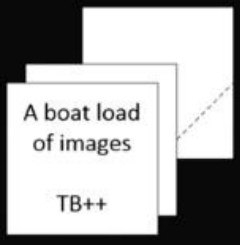
- Use unmanned
- Minminize risk to people
- 10x or more area
- Data to cloud
- Time to result in hours
- $$

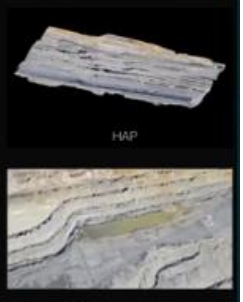
INSITU

aws

KEY ARCHITECTURE CHALLENGES

3





Days

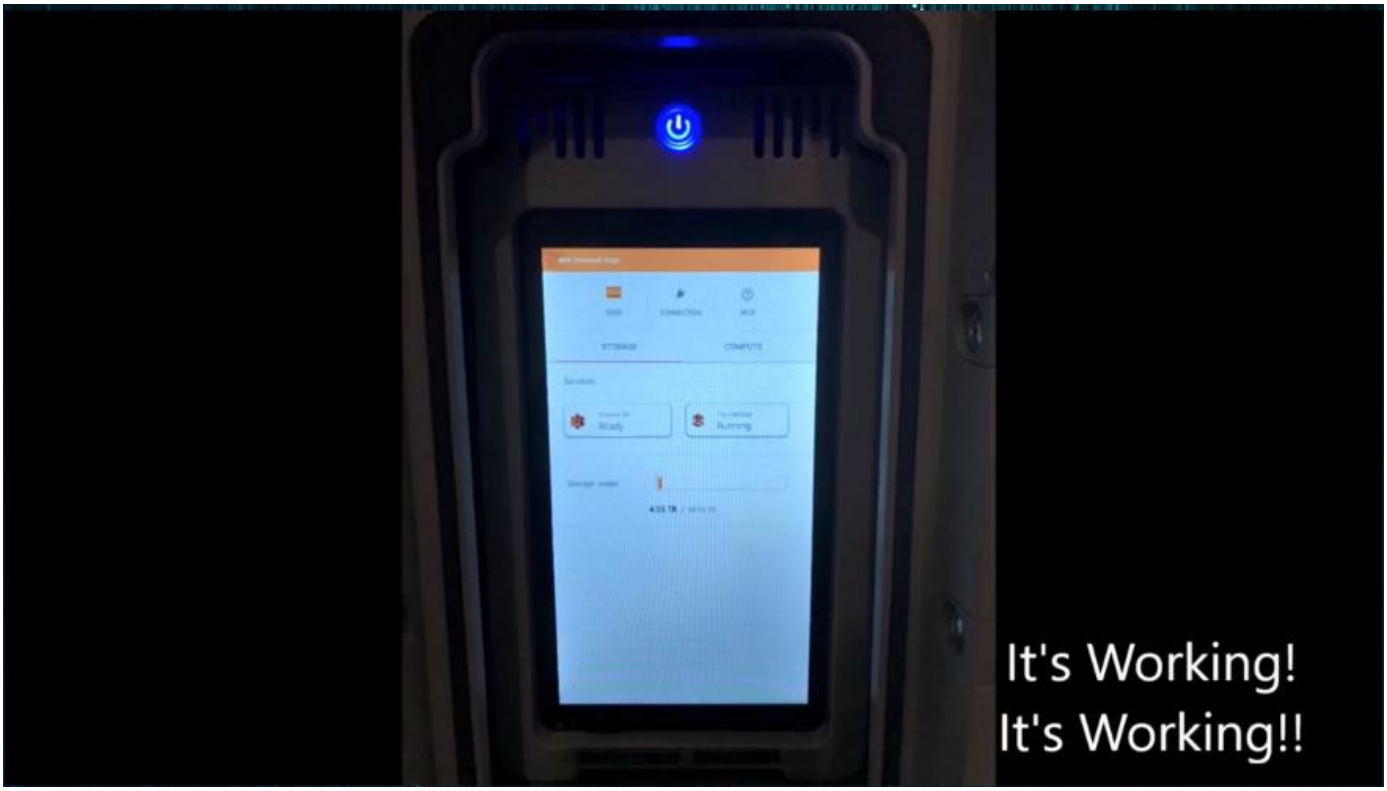
Finally we can start analysis

INEXA Cloud

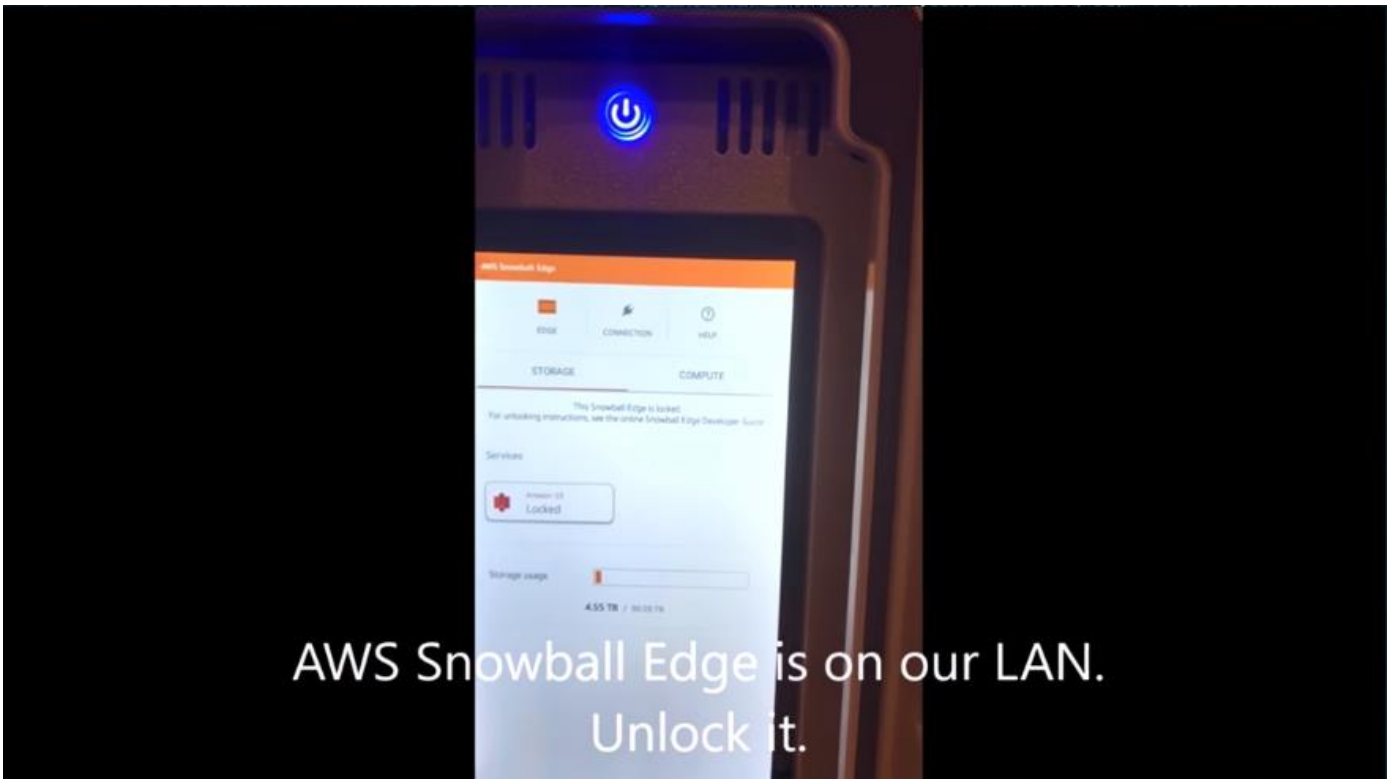
INSITU
INEXA CLOUD
AWS SNOWBALL EDGE
BG&O 2017

SH & RT

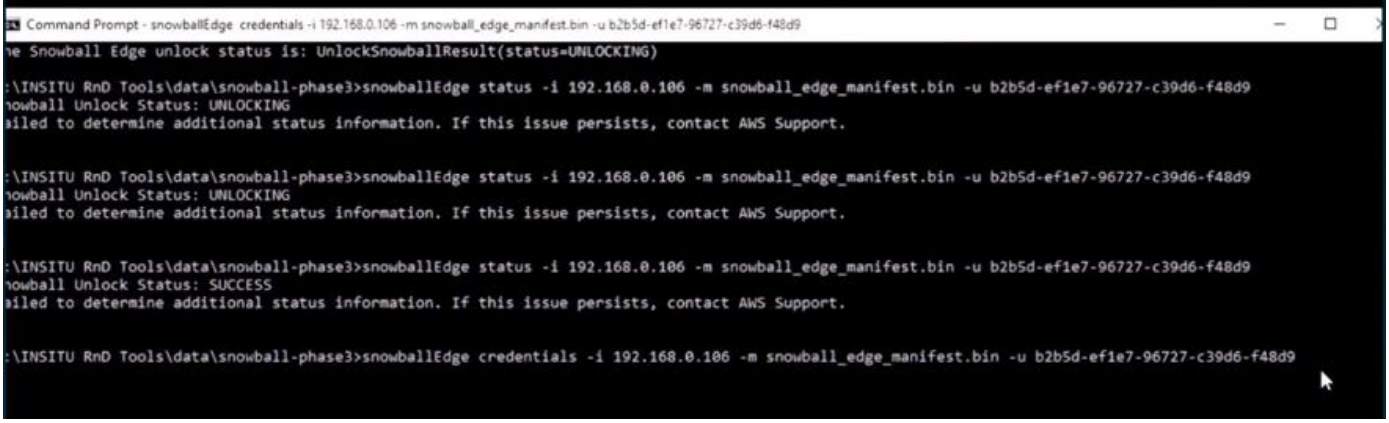

Start it up


It's Working!
It's Working!!

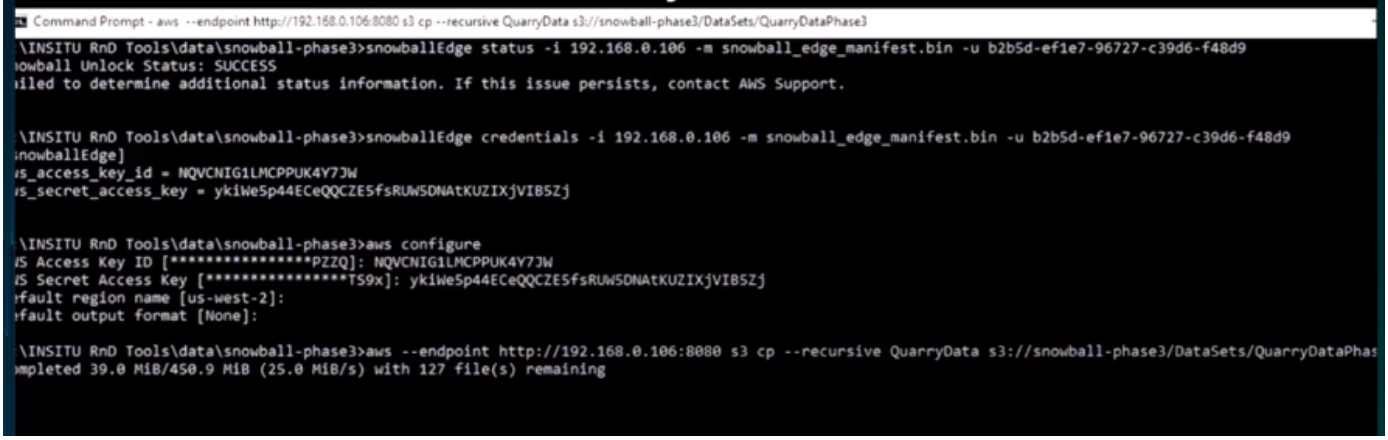

Command Prompt
* -i, --ip
The IP address of the Snowball Edge (or the leader node of
cluster). This can be found on the E Ink display.
* -m, --manifest
The path to the manifest file. This can be downloaded from the AWS
Console or retrieved from the job management API.
* -n, --nodeId
Use this option to provide the node id of the new node.
* -u, --unlockcode
The code used to unlock the appliance or cluster. This can be
downloaded from the AWS Console or retrieved from the job
management API.
:\INSITU RnD Tools\data\snowball-phase3>snowballEdge unlock -i 192.168.0.106 -m snowball_edge_manifest.bin -u b2b5d-ef1e7-96727-c39d6-f48d9
he Snowball Edge unlock status is: UnlockSnowballResult(status=UNLOCKING)
:\INSITU RnD Tools\data\snowball-phase3>
AWS Snowball Edge is on our LAN.
Unlock it.


STORAGE
COMPUTE
Locked
AWS Snowball Edge is on our LAN.
Unlock it.

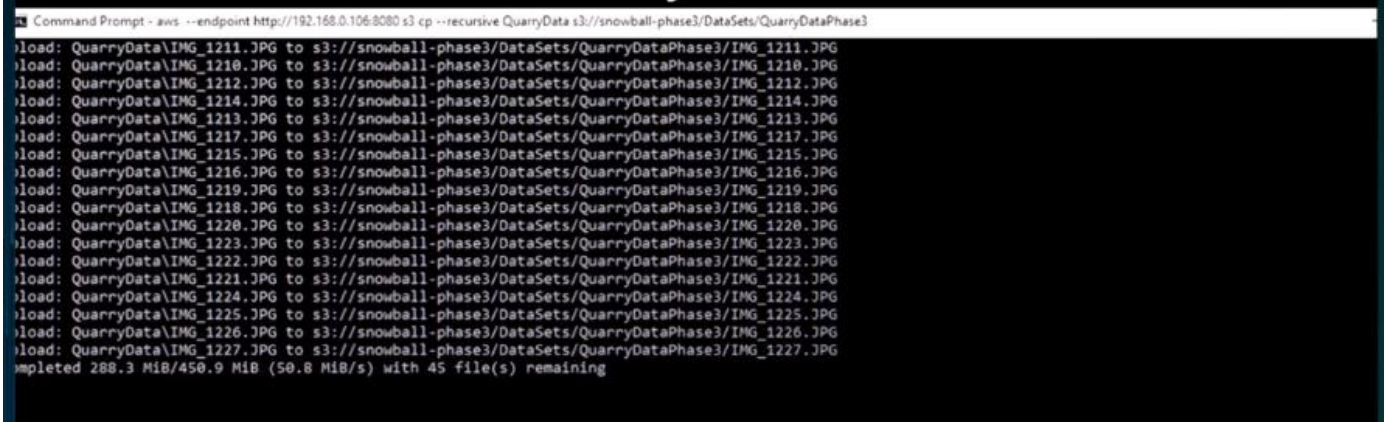
Authorize credentials

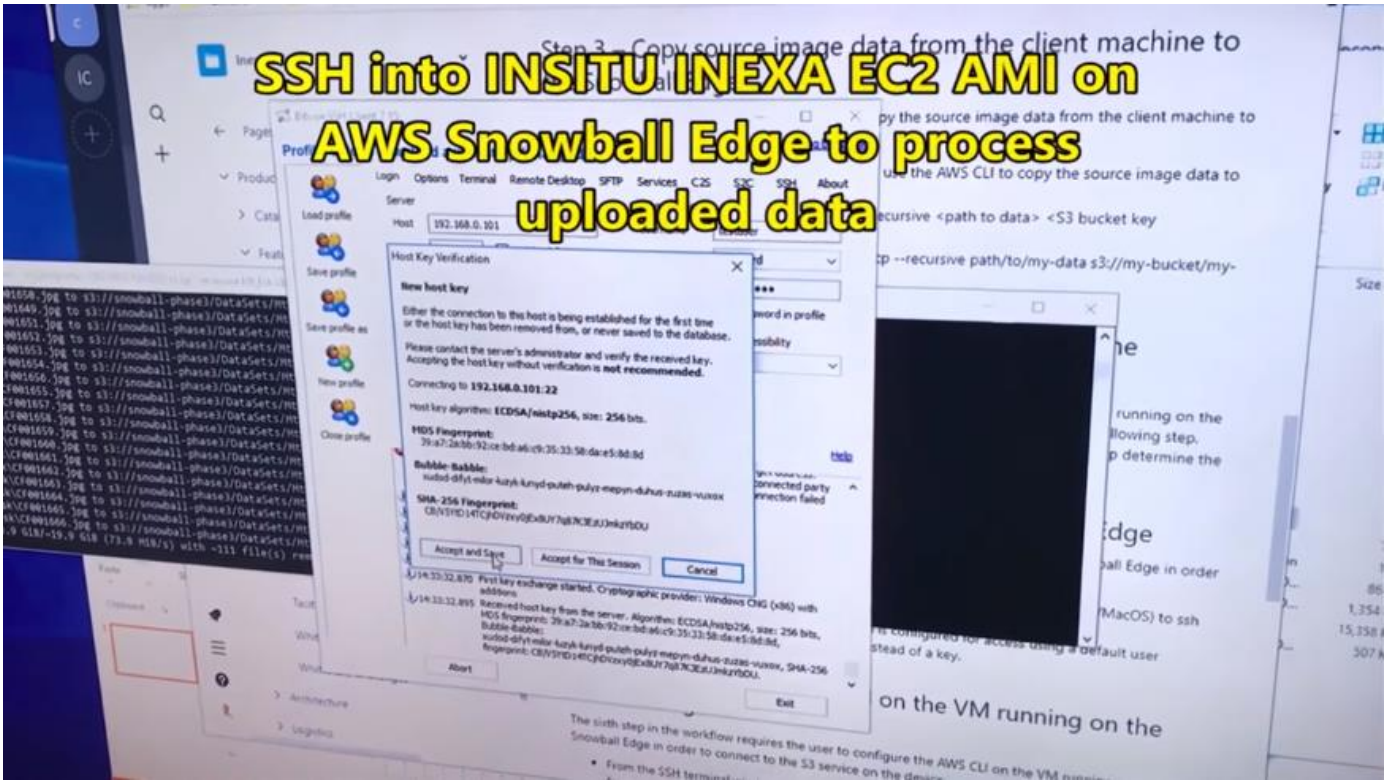


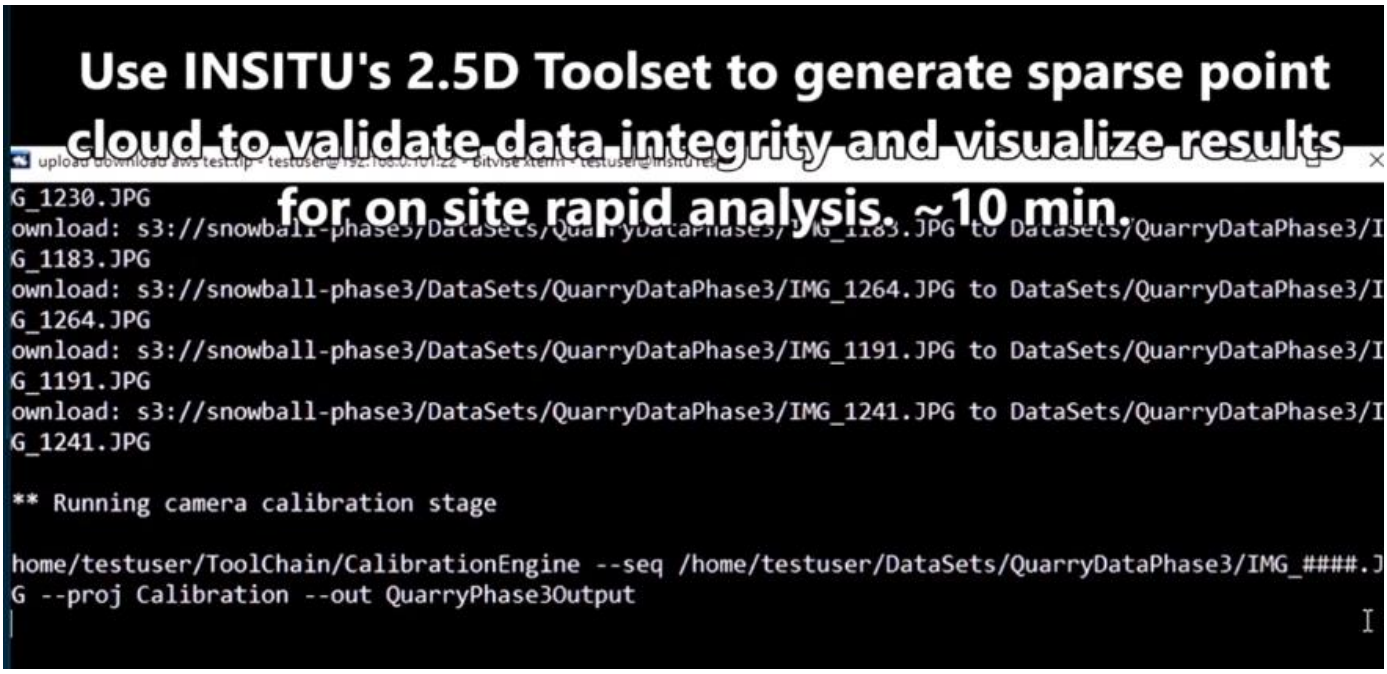
From local machine, into Snowball's S3 bucket, Upload data you will.

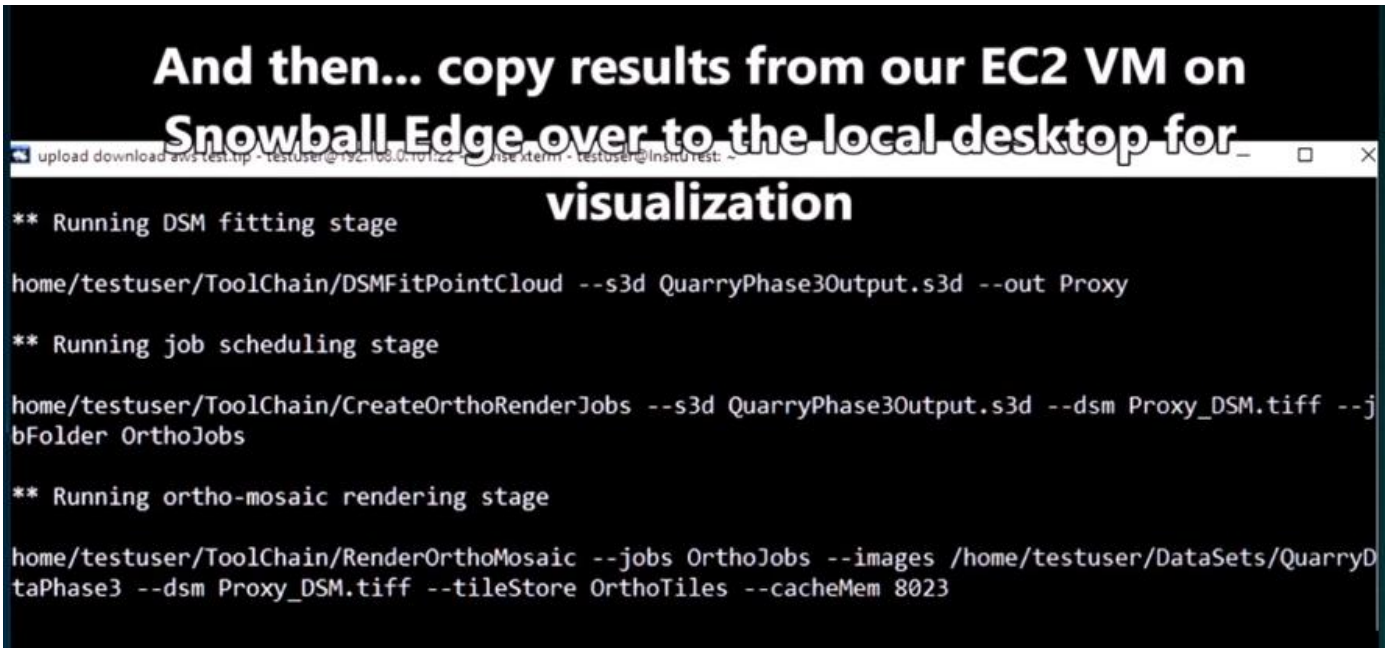


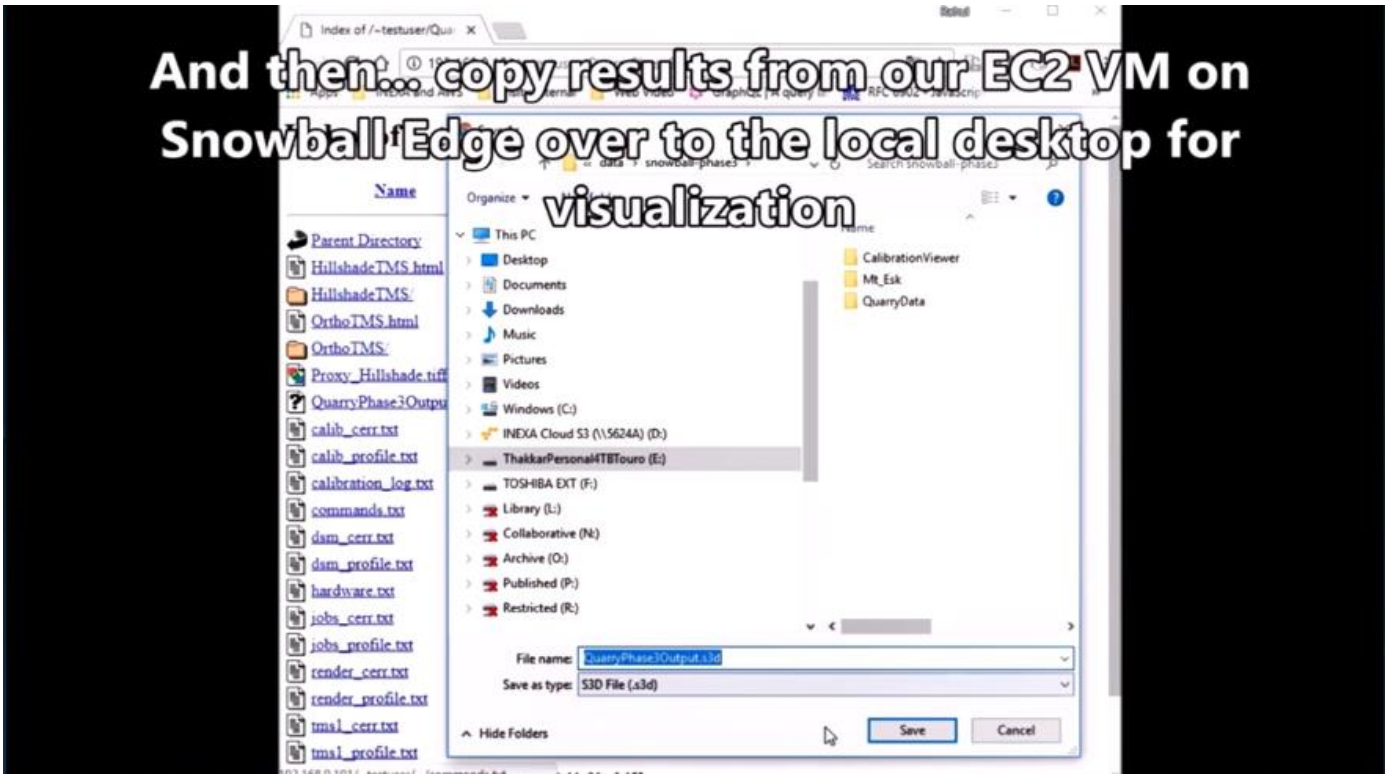
From local machine, into Snowball's S3 bucket, Upload data you will.

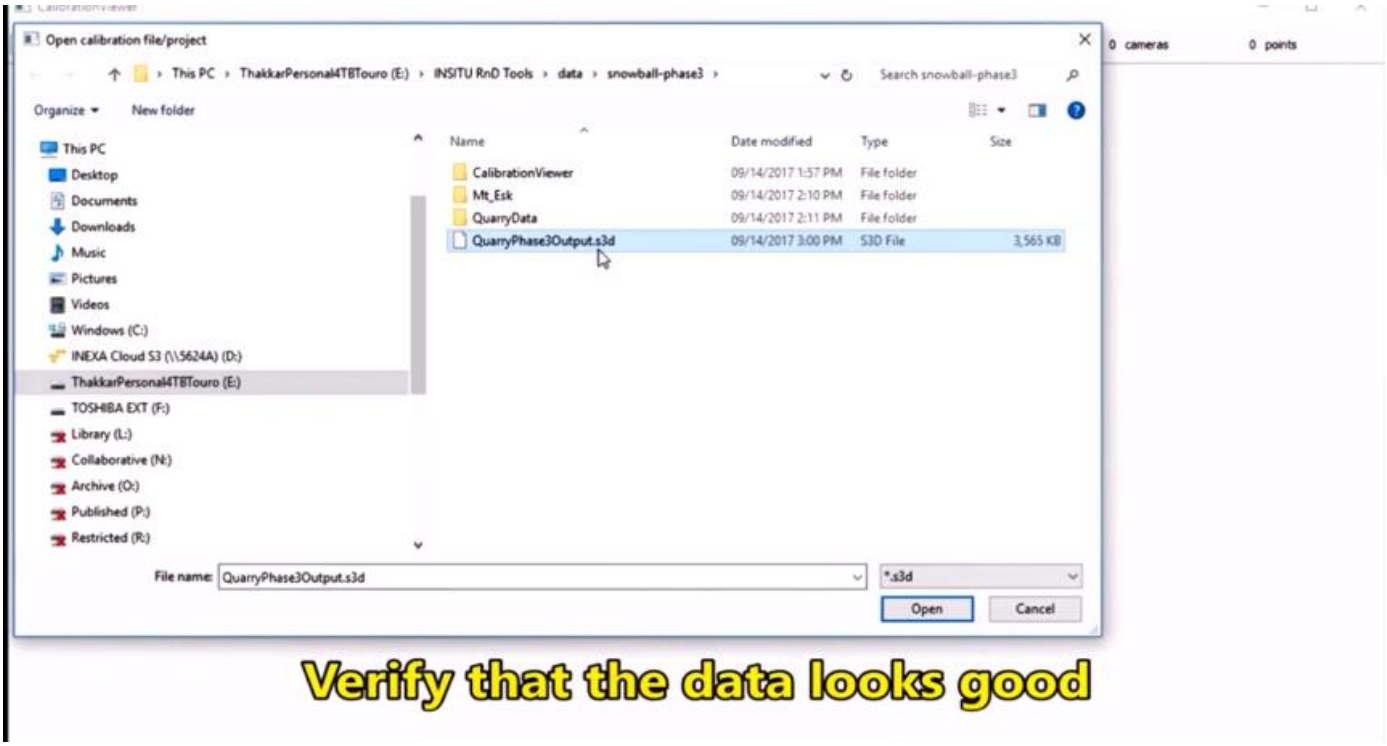



SSH into INSITU INEXA EC2 AMI on
AWS Snowball Edge to process
uploaded data


Use INSITU's 2.5D Toolset to generate sparse point
cloud to validate data integrity and visualize results
for on site rapid analysis. ~10 min.
G_1230.JPG
ownload: s3://snowball-phase3/DataSets/QuarryDataPhase3/IMG_1183.JPG to DataSets/QuarryDataPhase3/I
G_1183.JPG
ownload: s3://snowball-phase3/DataSets/QuarryDataPhase3/IMG_1264.JPG to DataSets/QuarryDataPhase3/I
G_1264.JPG
ownload: s3://snowball-phase3/DataSets/QuarryDataPhase3/IMG_1191.JPG to DataSets/QuarryDataPhase3/I
G_1191.JPG
ownload: s3://snowball-phase3/DataSets/QuarryDataPhase3/IMG_1241.JPG to DataSets/QuarryDataPhase3/I
G_1241.JPG
** Running camera calibration stage
home/testuser/ToolChain/CalibrationEngine --seq /home/testuser/DataSets/QuarryDataPhase3/IMG_####.J
G --proj Calibration --out QuarryPhase3Output

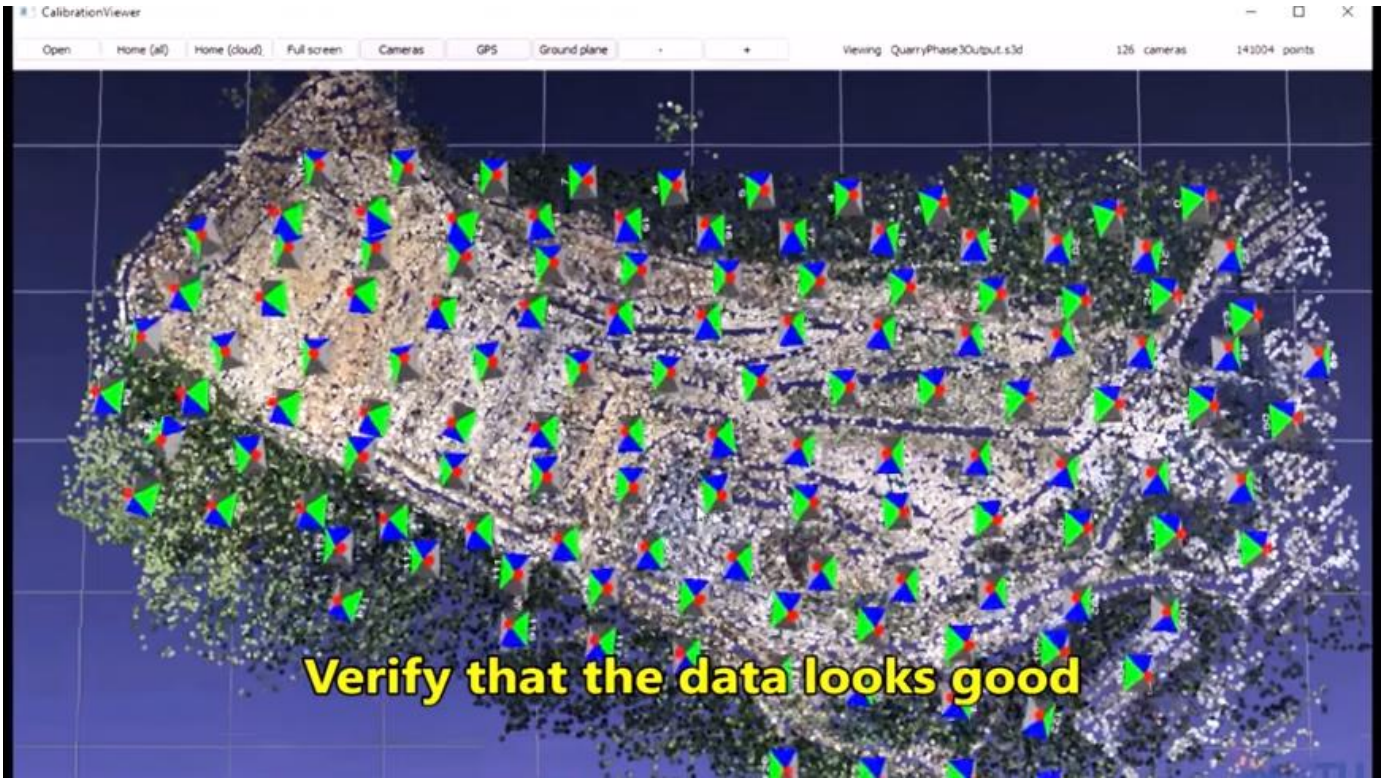

And then... copy results from our EC2 VM on Snowball Edge over to the local desktop for visualization
** Running DSM fitting stage
home/testuser/ToolChain/DSMFitPointCloud --s3d QuarryPhase3Output.s3d --out Proxy
** Running job scheduling stage
home/testuser/ToolChain/CreateOrthoRenderJobs --s3d QuarryPhase3Output.s3d --dsm Proxy_DSM.tiff --j
bFolder OrthoJobs
** Running ortho-mosaic rendering stage
home/testuser/ToolChain/RenderOrthoMosaic --jobs OrthoJobs --images /home/testuser/DataSets/QuarryD
taPhase3 --dsm Proxy_DSM.tiff --tileStore OrthoTiles --cacheMem 8023

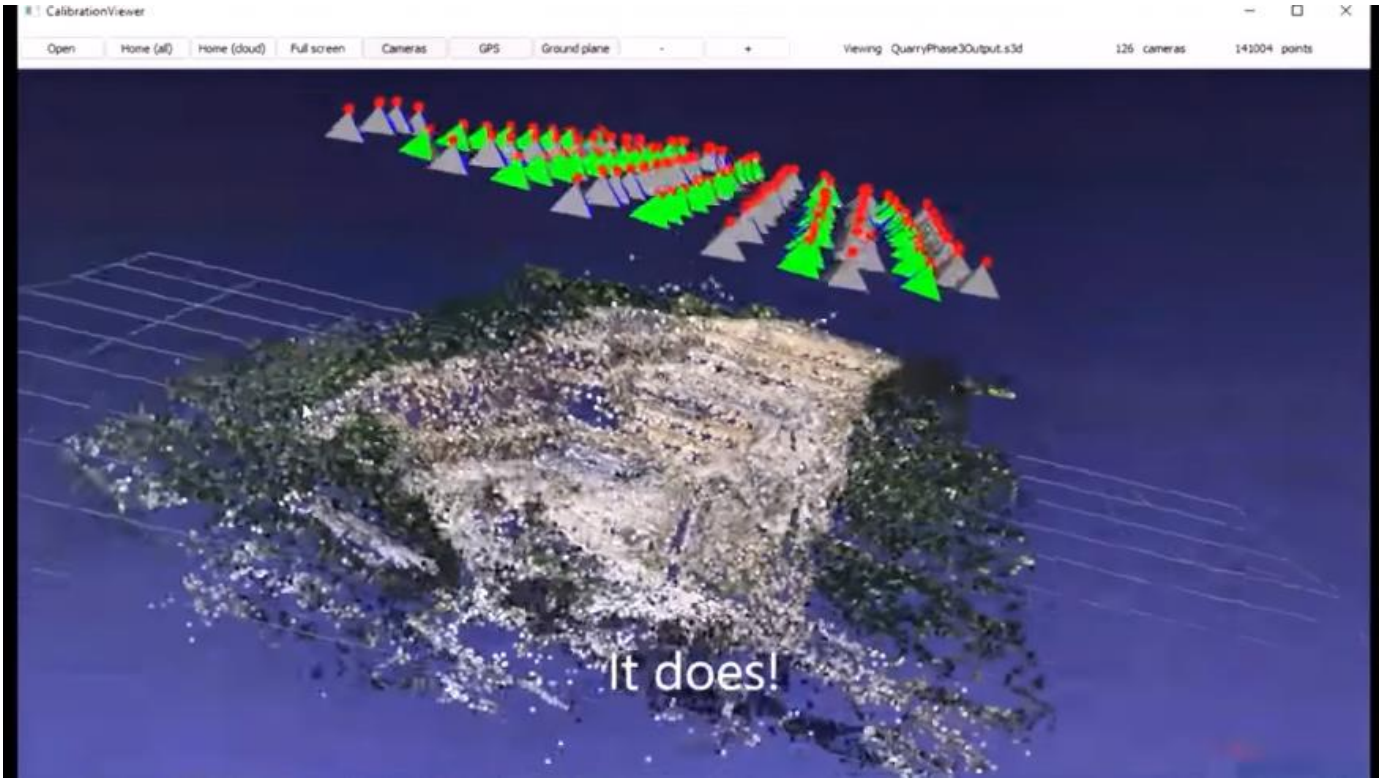

And then... copy results from our EC2 VM on Snowball Edge over to the local desktop for visualization
Name
Parent Directory
HillshadeTMS.html
HillshadeTMS/
OrthoTMS.html
OrthoTMS/
Proxy_Hillshade.tiff
QuarryPhase3Output
calib_cerr.txt
calib_profile.txt
calibration_log.txt
commands.txt
dsm_cerr.txt
dsm_profile.txt
hardware.txt
jobs_cerr.txt
jobs_profile.txt
render_cerr.txt
render_profile.txt
tms1_cerr.txt
tms1_profile.txt
Organize
This PC
Desktop
Documents
Downloads
Music
Pictures
Videos
Windows (C:)
INEXA Cloud S3 (\\5624A) (D:)
ThakkarPersonal4TBTouro (E:)
TOSHIBA EXT (F:)
Library (L:)
Collaborative (N:)
Archive (O:)
Published (P:)
Restricted (R:)
CalibrationViewer
Mt_Esk
QuarryData
File name: QuarryPhase3Output.s3d
Save as type: S3D File (.s3d)
Hide Folders
Save
Cancel



Loading calibration

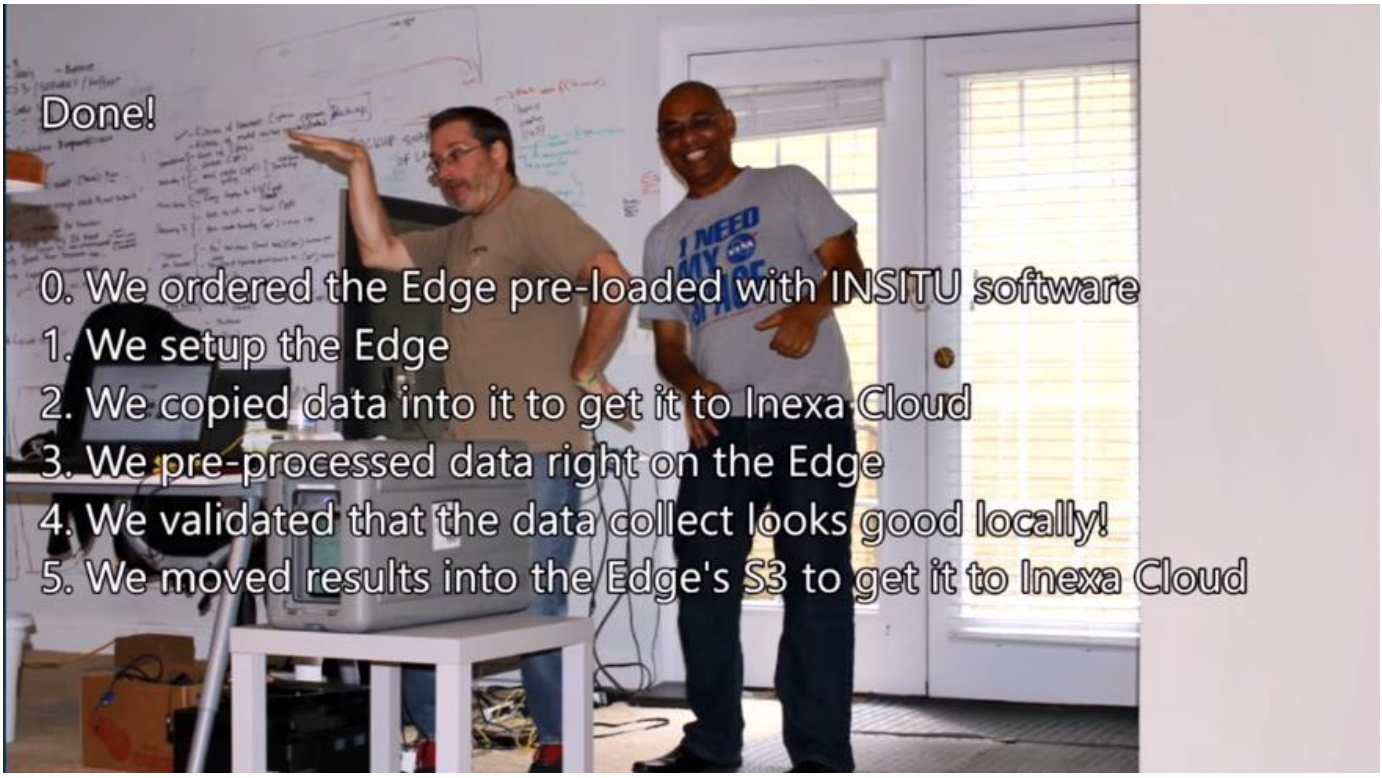
Verify that the data looks good


CalibrationViewer
Open
Home (all)
Home (cloud)
Full screen
Cameras
GPS
Ground plane
-
+
Viewing QuarryPhase3Output.s3d
126 cameras
141004 points
Verify that the data looks good


CalibrationViewer
Open
Home (all)
Home (cloud)
Full screen
Cameras
GPS
Ground plane
-
+
Viewing QuarryPhase3Output.s3d
126 cameras
141004 points
It does!

From the INSITU EC2 instance on the Snowball, transfer the data into Snowball's S3 bucket

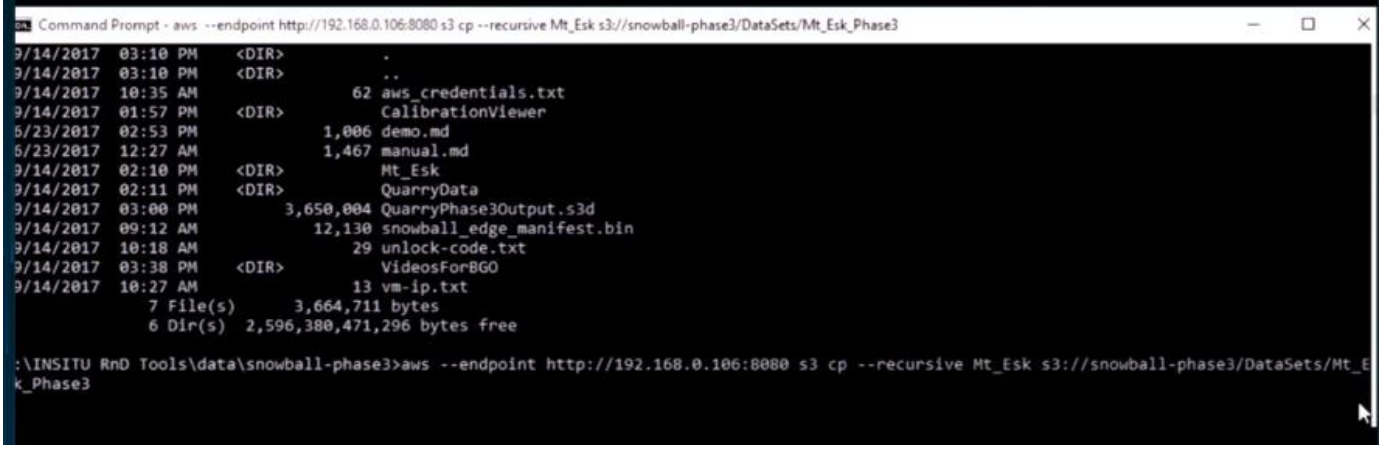
```
upload download aws test.tip - testuser@192.168.0.101:22 - Bitvise xterm - testuser@InsituTest: ~/public_html
pload: QuarryPhase3Output/OrthoTMS/18/271666/199014.png to s3://snowball-phase3/DataSets/QuarryPhas
3Output/OrthoTMS/18/271666/199014.png
pload: QuarryPhase3Output/OrthoTMS/18/271666/199019.png to s3://snowball-phase3/DataSets/QuarryPhas
3Output/OrthoTMS/18/271666/199019.png
pload: QuarryPhase3Output/OrthoTMS/18/271666/199017.png to s3://snowball-phase3/DataSets/QuarryPhas
3Output/OrthoTMS/18/271666/199017.png
pload: QuarryPhase3Output/OrthoTMS/18/271666/199018.png to s3://snowball-phase3/DataSets/QuarryPhas
3Output/OrthoTMS/18/271666/199018.png
pload: QuarryPhase3Output/OrthoTMS/18/271667/199013.png to s3://snowball-phase3/DataSets/QuarryPhas
3Output/OrthoTMS/18/271667/199013.png
pload: QuarryPhase3Output/OrthoTMS/18/271667/199014.png to s3://snowball-phase3/DataSets/QuarryPhas
3Output/OrthoTMS/18/271667/199014.png
pload: QuarryPhase3Output/OrthoTMS/18/271667/199018.png to s3://snowball-phase3/DataSets/QuarryPhas
3Output/OrthoTMS/18/271667/199018.png
ompleted 16.3 MiB/~17.9 MiB with ~12 file(s) remaining (calculating...)
```



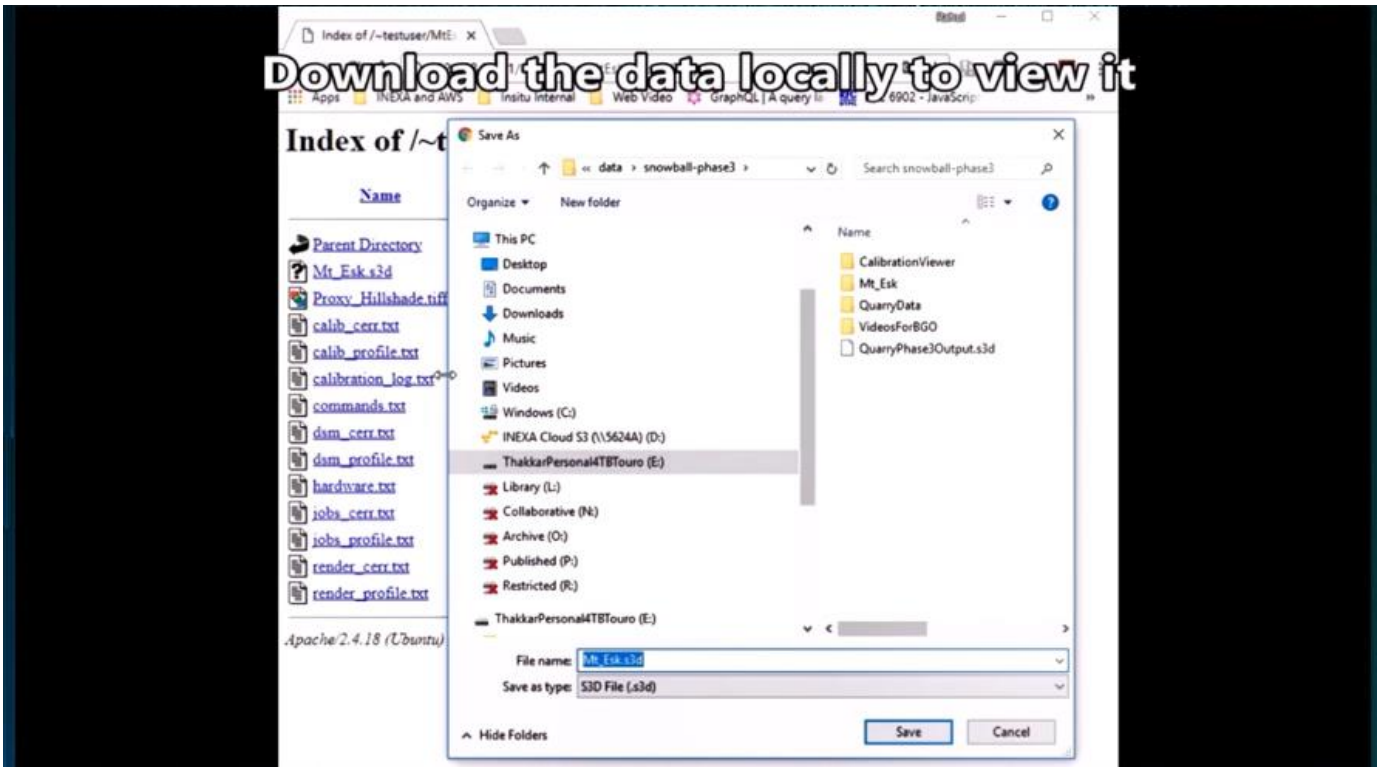
But wait!!!

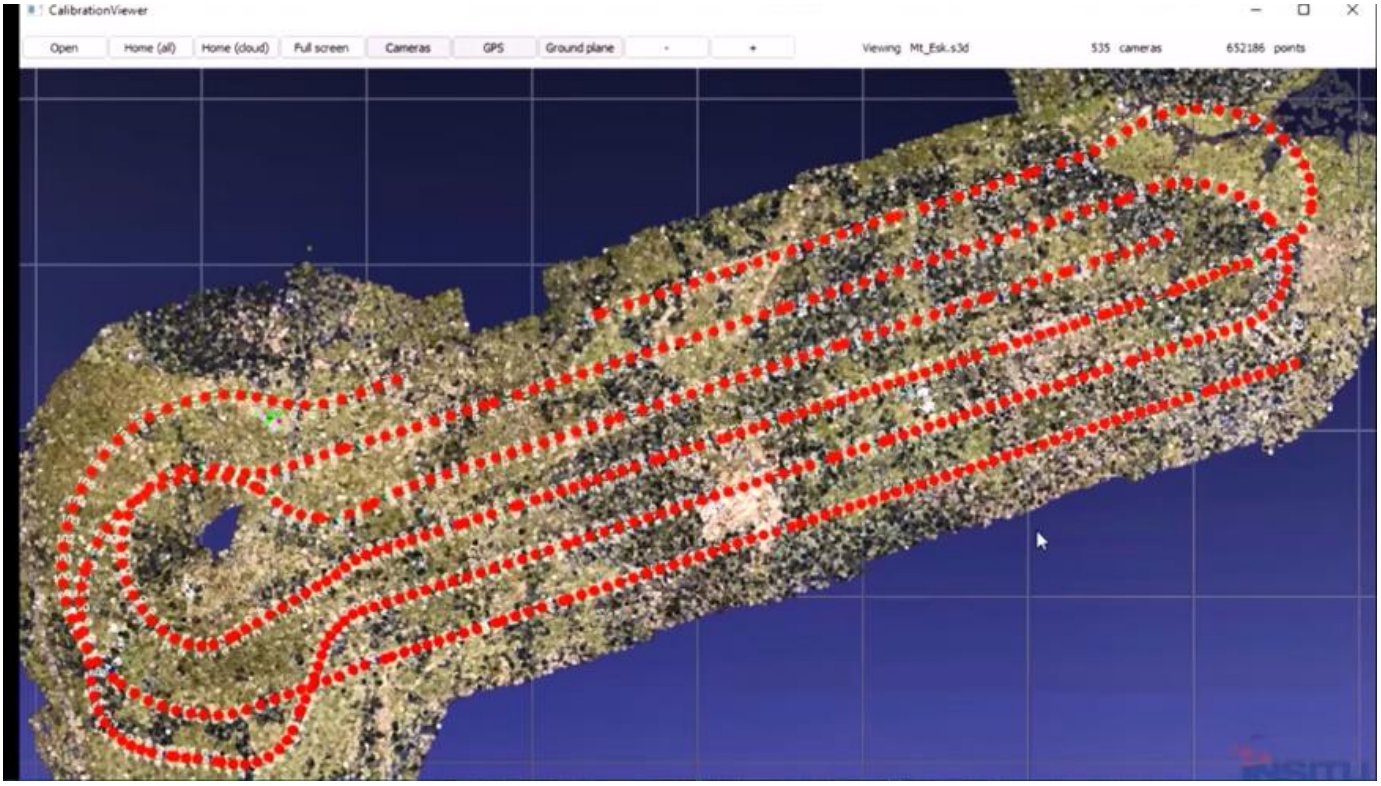
What if we have more data?

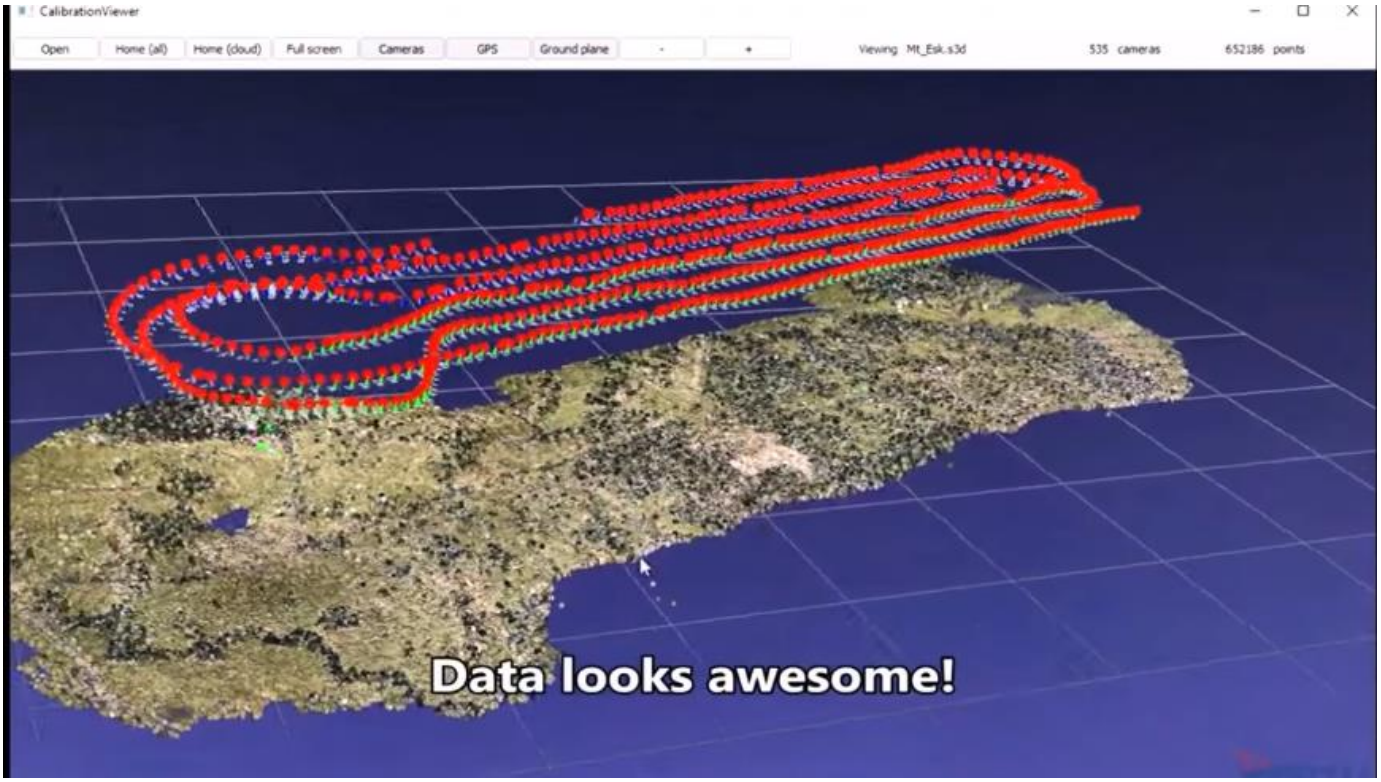
Well... let's upload it into the Snowball's S3 for processing...

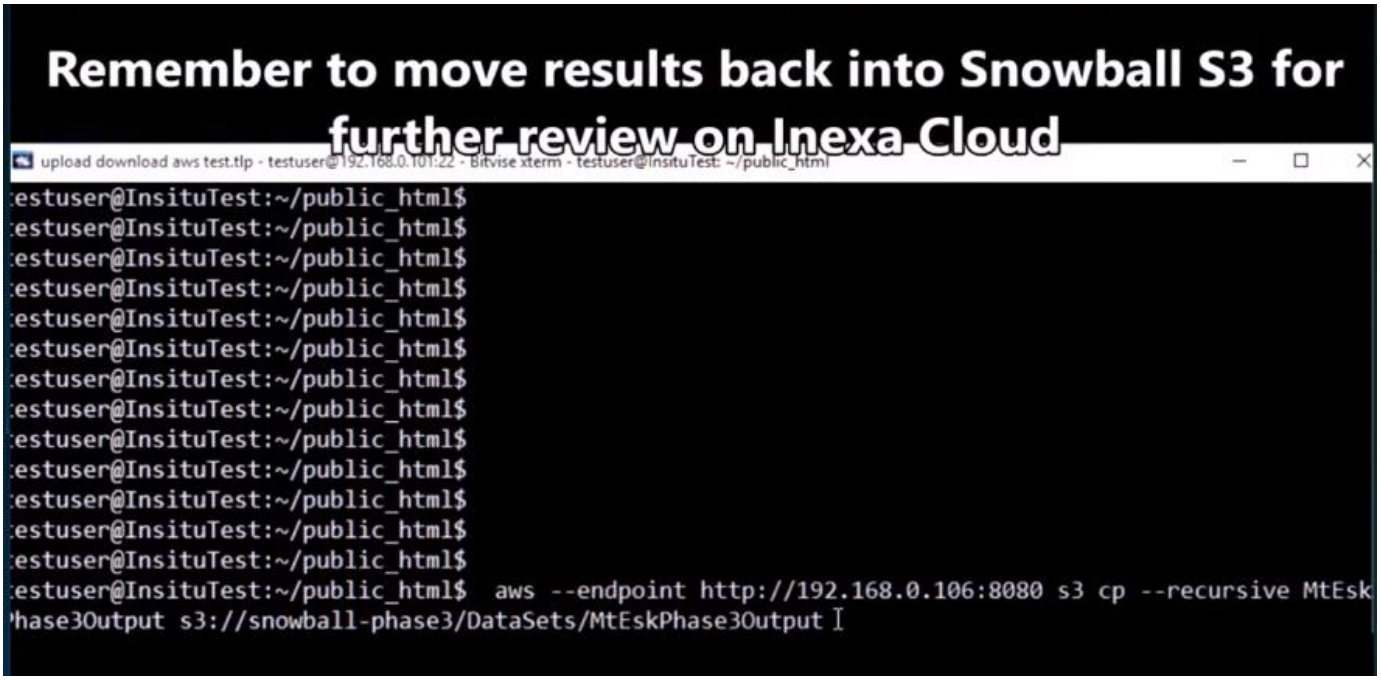


This is about 30 GB Mt. Esk data collected near Brisbane Australia using the HAP payload being processed in an hour or so using INSITU software pre-loaded on the Edge before it was shipped...


Download the data locally to view it
Index of /~t
Name
Parent Directory
Mt_Esk.s3d
Proxy_Hillshade.tiff
calib_cerr.txt
calib_profile.txt
calibration_log.txt
commands.txt
dsm_cerr.txt
dsm_profile.txt
hardware.txt
jobs_cerr.txt
jobs_profile.txt
render_cerr.txt
render_profile.txt
Apache/2.4.18 (Ubuntu)
Save As
data › snowball-phase3 ›
Search snowball-phase3
Organize
New folder
This PC
Desktop
Documents
Downloads
Music
Pictures
Videos
Windows (C:)
INEXA Cloud S3 (\\5624A) (D:)
ThakkarPersonal4TBTouro (E:)
Library (L:)
Collaborative (N:)
Archive (O:)
Published (P:)
Restricted (R:)
Name
CalibrationViewer
Mt_Esk
QuarryData
VideosForBGO
QuarryPhase3Output.s3d
File name: Mt_Esk.s3d
Save as type: S3D File (.s3d)
Hide Folders
Save
Cancel


CalibrationViewer
Open
Home (all)
Home (cloud)
Full screen
Cameras
GPS
Ground plane
Viewing Mt_Esk.s3d
535 cameras
652186 points


CalibrationViewer
Open
Home (all)
Home (cloud)
Full screen
Cameras
GPS
Ground plane
Viewing Mt_Esk.s3d
535 cameras
652186 points
Data looks awesome!

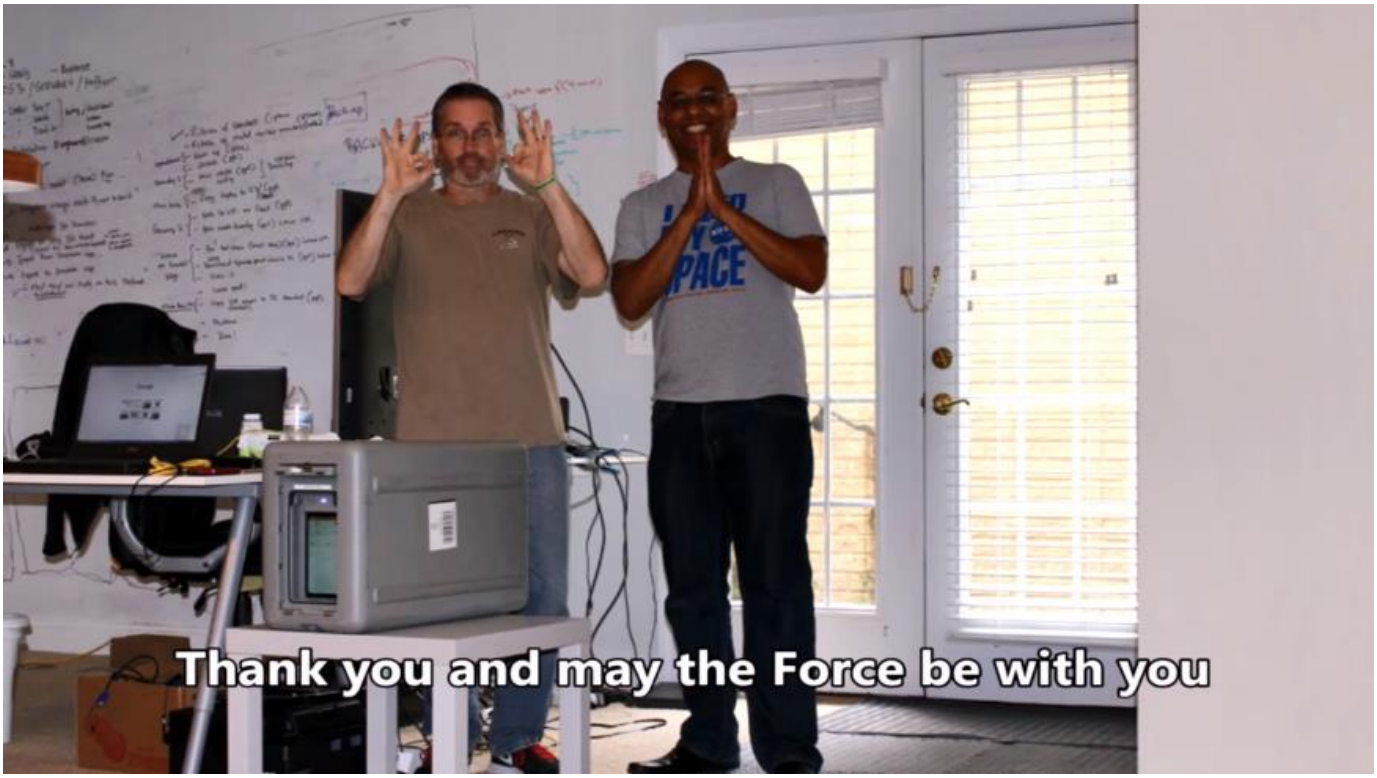

Remember to move results back into Snowball S3 for further review on Inexa Cloud
upload download aws test.tlp - testuser@192.168.0.101:22 - Bitvise xterm - testuser@InsituTest: ~/public_html
estuser@InsituTest:~/public_html$
estuser@InsituTest:~/public_html$
estuser@InsituTest:~/public_html$
estuser@InsituTest:~/public_html$
estuser@InsituTest:~/public_html$
estuser@InsituTest:~/public_html$
estuser@InsituTest:~/public_html$
estuser@InsituTest:~/public_html$
estuser@InsituTest:~/public_html$
estuser@InsituTest:~/public_html$
estuser@InsituTest:~/public_html$
estuser@InsituTest:~/public_html$
estuser@InsituTest:~/public_html$ aws --endpoint http://192.168.0.106:8080 s3 cp --recursive MtEsk
hase3Output s3://snowball-phase3/DataSets/MtEskPhase3Output

Remember to move results back into Snowball S3 for further review on Inexa Cloud

upload download aws test.tlp - testuser@192.168.0.101:22 - Bitvise xterm - testuser@InsituTest: ~/public_html

```
:t
pload: MtEskPhase3Output/render_profile.txt to s3://snowball-phase3/DataSets/MtEskPhase3Output/rend
:r_profile.txt
pload: MtEskPhase3Output/jobs_profile.txt to s3://snowball-phase3/DataSets/MtEskPhase3Output/jobs_p
:ofile.txt
pload: MtEskPhase3Output/render_cerr.txt to s3://snowball-phase3/DataSets/MtEskPhase3Output/render_
:err.txt
pload: MtEskPhase3Output/jobs_cerr.txt to s3://snowball-phase3/DataSets/MtEskPhase3Output/jobs_cerr
txt
pload: MtEskPhase3Output/calibration_log.txt to s3://snowball-phase3/DataSets/MtEskPhase3Output/cal
.bration_log.txt
pload: MtEskPhase3Output/Proxy_Hillshade.tiff to s3://snowball-phase3/DataSets/MtEskPhase3Output/Pr
)xy_Hillshade.tiff
pload: MtEskPhase3Output/Mt_Esk.s3d to s3://snowball-phase3/DataSets/MtEskPhase3Output/Mt_Esk.s3d
:estuser@InsituTest:~/public_html$
```

Power it down and ship it... Soon your data will be on Inexa Cloud



New Storage Training

aws training and certification

For Enterprise Storage Engineers

- Learn how to architect and manage highly available solutions on AWS storage services
- Advance toward AWS certifications
- Help your organization migrate to the cloud faster

Online at www.aws.training

- Access 100+ new digital training courses including advanced training on storage
- Deep Dives on S3, EFS, and EBS
- Migrating and Tiering Storage to AWS (Hybrid Solutions)

At re:Invent

- Visit Hands-on Labs at the Venetian
- Attend a proctored "Introduction to EFS" Spotlight Lab on Thursday at 3pm at the Venetian
- Meet Storage experts at the Ask the Experts in Hands-on Labs room at the Venetian

AWS re:Invent

aws