Coping With Aging (Data)

Keith Gregory
AWS Practice Lead
Chariot Solutions

IoT on AWS

A Philly Cloud Computing Event

CHARIOT SOLUTIONS

Data Has Different Roles As It Ages

Young (0-4 hours)	Alerts & Feedback
Middle-Aged (0-4 weeks)	Analytics & Reporting
Elderly (0-forever)	Model development and testing Auditing

CHARIOT SOLUTIONS

Storage and Access Should Change Too

Young (0-4 hours)	Direct processing of input stream Core IoT rule
Middle-Aged (0-4 weeks)	Database Data lake
Elderly (0-forever)	Partitioned data lake

CHARIOT SOLUTIONS

Databases?

Relational	Allows joins between IoT and auxiliary data
Time Series	Optimized for data that is organized by time (particularly increasing time) Typically provide query enhancements, such as gap-filling or sampling May not support joins, often restrict ability to index data
Search Engine	Fast queries based on metadata
DynamoDB	You probably don't want to use, unless you're retrieving by primary key

CHARIOT SOLUTIONS

Data Lakes Exploit Low Cost Storage and CPU

A (large) repository of simple data files …

That can be processed in parallel …

Without indexes …

CHARIOT SOLUTIONS

Data Lakes are not Oceans … or Swamps

Organize your data into multiple lakes

Raw, Cooked, Auxiliary, …

If it has a different schema, it should be in a different lake

Use a “lake friendly” format for analysis

Columnar data formats (eg, Parquet) are usually smaller/faster

If using text formats, always GZip!

Acquire Auxiliary Data with IoT Data

Example: outside temperature

Historic data may not be available!

Can be challenging to wire data together

This is where relational DB is great

Transformation step as data is ingested

CHARIOT SOLUTIONS

Retain Everything, Unless Prevented

Include stuff you made up

It's usually easier (and cheaper!) to store rather than re-create

"Active" data just is a view on top of your real data

It might have a different format, or filtering

Sometimes, regulations get in the way

GPDR, California Consumer Privacy Act (CCPA), …

Corporate records-retention policies

CHARIOT SOLUTIONS

Whaddaya mean “partitioned” data lake?

“2013 called, they’re wondering why you’re reading their data”

Unless you need the data, you don’t want to pay to read it.

A partitioned data lake uses “folders” to segment data

Remember that S3 doesn’t really have folders, or file globbing

Athena uses a path convention to define partitioning columns

CHARIOT SOLUTIONS

Plan for Change

Example: new sensor reads slightly lower than old

- Do you compensate all readings?
- Update your old data?
- Use different calculations depending on which sensor supplied the data?

Retain metadata (eg, sensor model)

Sensors may be repurposed!

CHARIOT SOLUTIONS

References

Amazon Athena User Guide
https://docs.aws.amazon.com/athena/latest/ug/what-is.html

Amazon Redshift Developer Guide
https://docs.aws.amazon.com/redshift/latest/dg/welcome.html

Elasticsearch as a Time Series Data Store
https://www.elastic.co/blog/elasticsearch-as-a-time-series-data-store

InfluxDB 1.7 documentation
https://docs.influxdata.com/influxdb/v1.7/

CHARIOT SOLUTIONS