PERCONA LIVE ONLINE MAY 12 - 13th 2021

Wide Partitions Data Modeling

Hello!

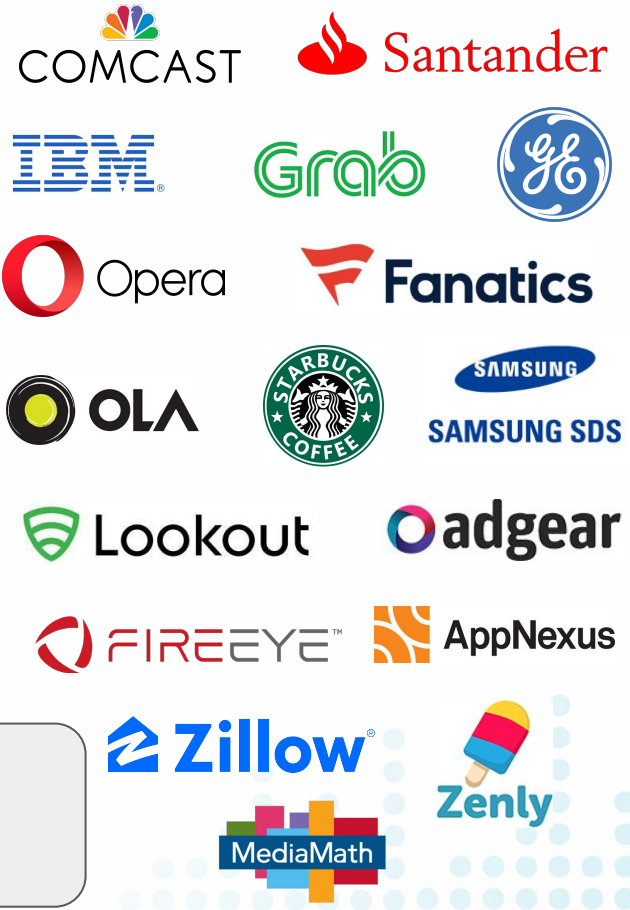
I am Tzach Livyatan

I know something about NoSQL

You can find me at tzach@scylladb.com

About ScyllaDB

- The Real-Time Big Data Database
- Drop-in replacement for Apache Cassandra and Amazon DynamoDB
- 10X the performance & low tail latency
- Open Source, Enterprise and Cloud options
- Founded by the creators of KVM hypervisor
- HQs: Palo Alto, CA, USA; Herzelia, Israel; Warsaw, Poland



Agenda

NoSQL Intro

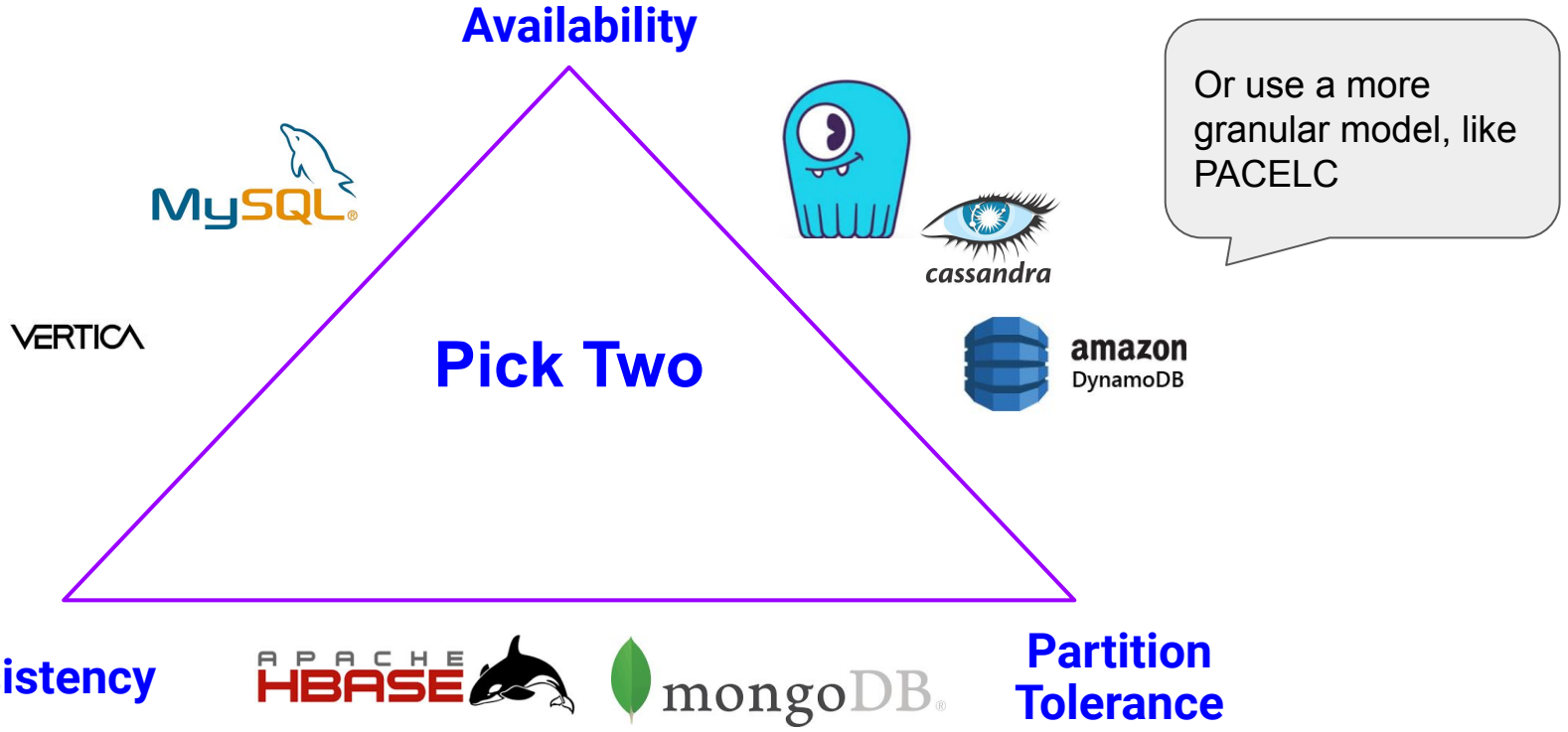
Wide Partition Data Modeling

- Choosing the right Partition Key
- Choosing the right Clustering Key
- Using Materialized Views / Secondary Index

Summary

1.
NoSQL in 5 mins

NoSQL - By Availability vs Consistency



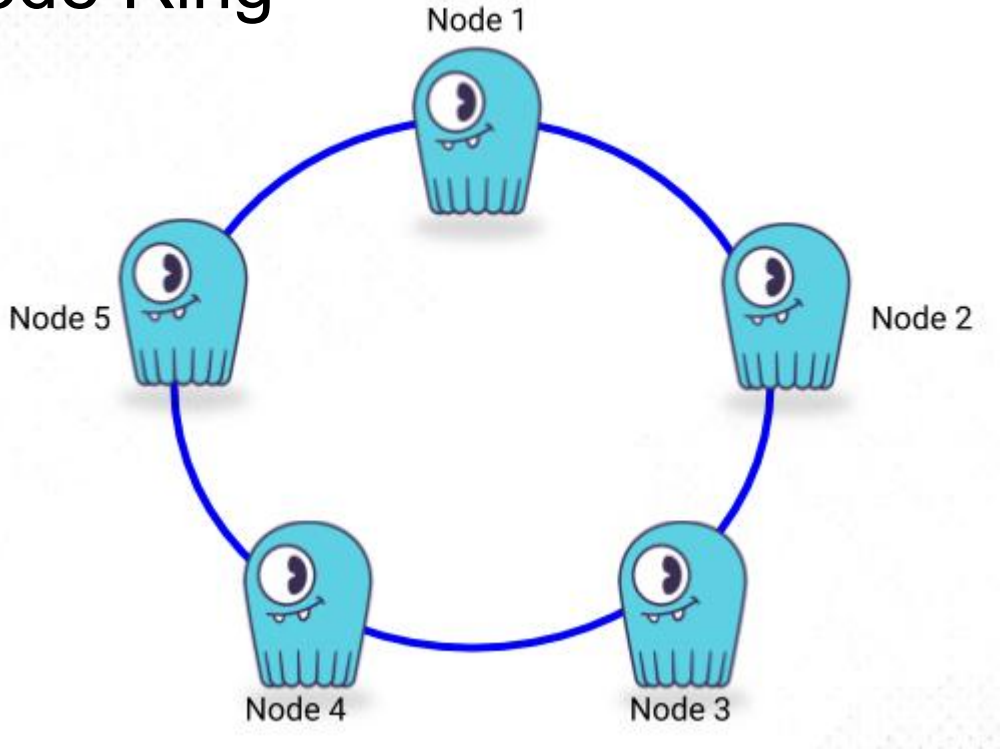
Or use a more granular model, like PACELC

NoSQL - By Data Model

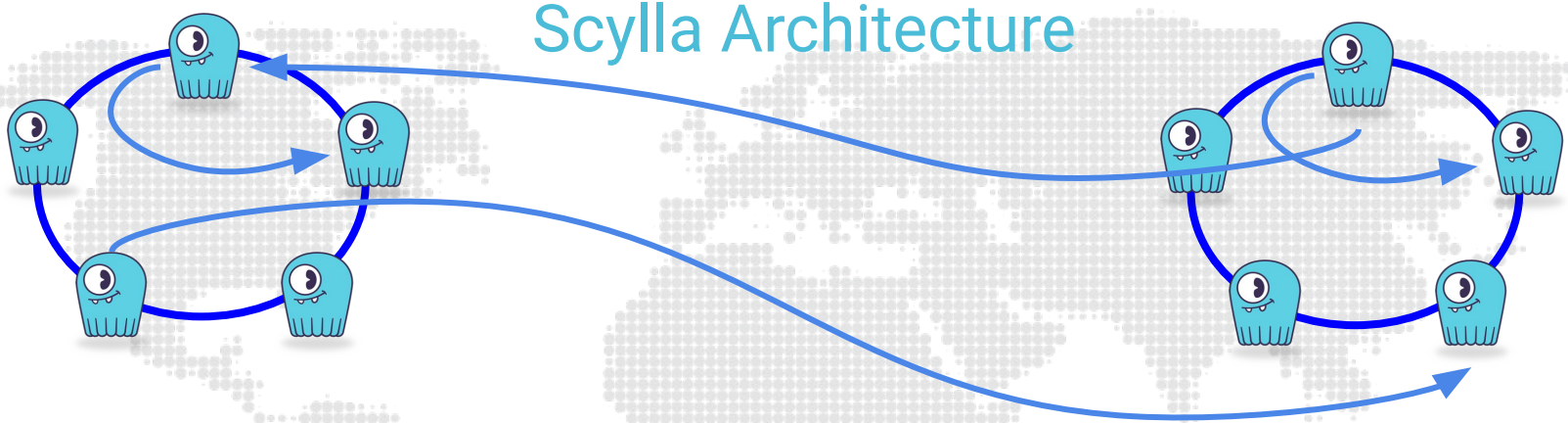
Complexity

Key / Value	 RocksDB AEROSPIKE  redis
Document store	 Couchbase
Wide column store	   amazon DynamoDB APACHE HBASE 
Graph	 JanusGraph 

Cluster - Node Ring



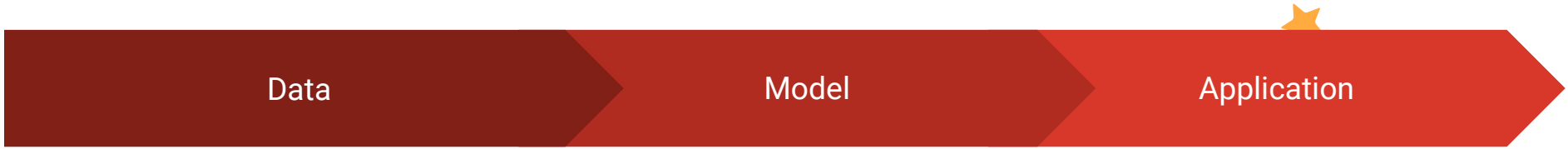
Scylla Architecture



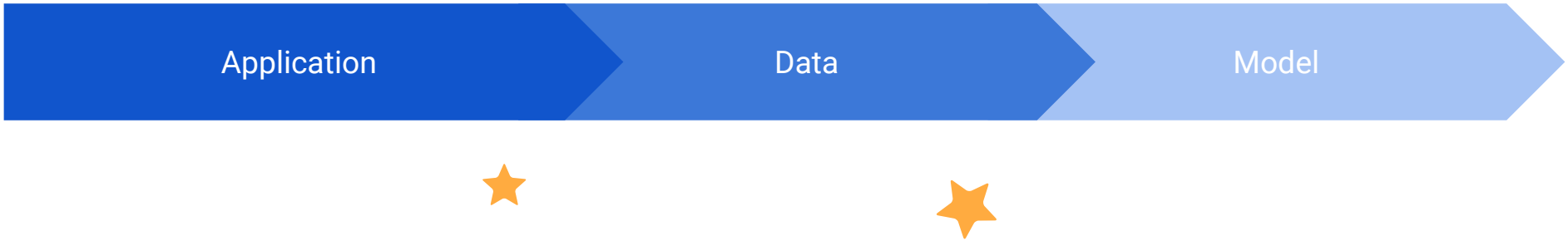
Active/active, replicated, auto-sharded

2.
Data Modeling

Relational:



NoSQL:



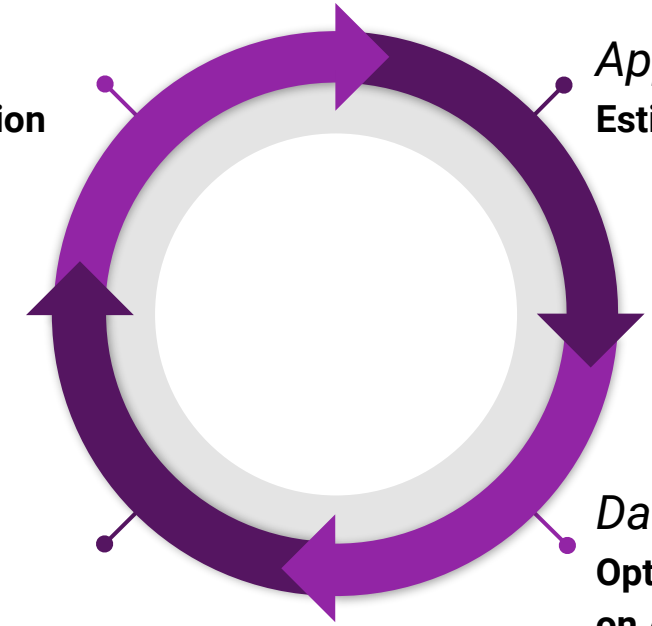
Circle of (NoSQL) Life

Update

Update schema / application

Application

Estimate Read and Write pattern



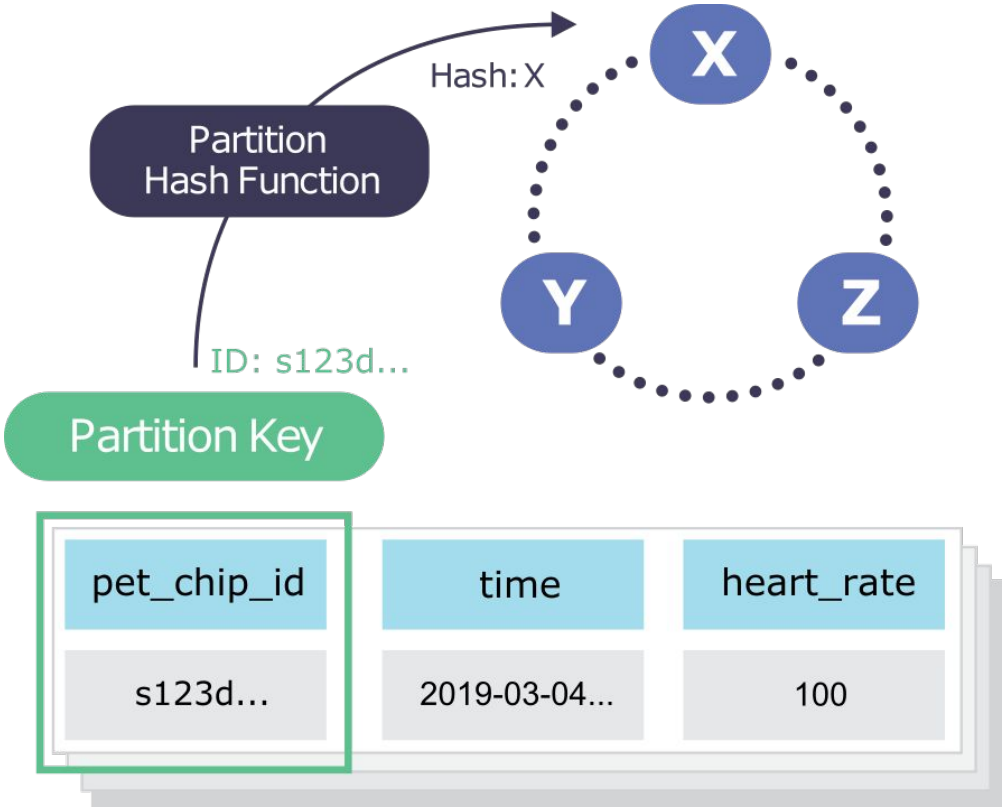
Measure

Use Metric to evaluate the data model

Data Model

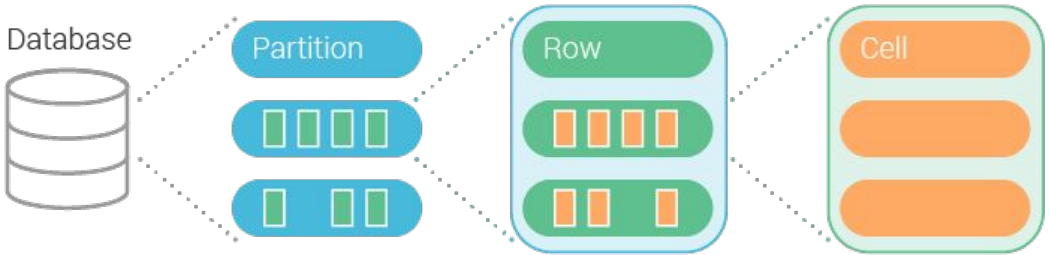
Optimize the model base on Apps SLA

Partition Key



Structure of data in Scylla

- Data is stored in MemTables and SSTables
- The highest level are **partitions**, identified by a partition key
- Partitions contain **rows**, identified by a clustering key
- Data is sorted in the partition by the clustering key, helping the storage engine to store data efficiently for querying



Key / Value Example

```
CREATE TABLE pet_owner (
    pet_chip_id uuid,
    owner uuid,
    pet_name text,
    PRIMARY KEY (pet_chip_id)
);
```

Partition Key

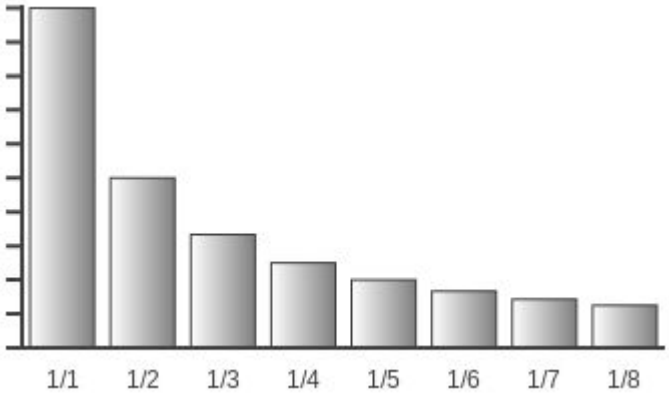
pet_chip_id	owner	pet_name
80d39c78-9dc0-11eb-a8b3-0242ac130003	642adfee-6ad9-...	Buddy
80d39c78-9dc0-11eb-a8b3-0242ac130003	642adfee-6ad9-...	Rocky
80d39c78-9dc0-11eb-a8b3-0242ac130003	642adfee-6ad9-...	Cat
...	...	...

Choosing a Partition Key

- High Cardinality
- Even Distribution

Avoid

- Low Cardinality
- Hot Partition
- Large Partition



https://www.codedrome.com/zipfs-law-in-python/

Key / Value Example

```
CREATE TABLE pet_owner (
    pet_chip_id uuid,
    owner uuid,
    pet_name text,
    PRIMARY KEY (pet_chip_id)
);
```

Partition Key

pet_chip_id	owner	pet_name
80d39c78-9dc0-11eb-a8b3-0242ac130003	642adfee-6ad9-...	Buddy
80d39c78-9dc0-11eb-a8b3-0242ac130003	642adfee-6ad9-...	Rocky
80d39c78-9dc0-11eb-a8b3-0242ac130003	642adfee-6ad9-...	Cat
...	...	...


Partition Key (pet_chip_id)
Partition (Rows)
268e074a-a801-476c-8db5-276eb2283b03
owner
pet_name
642adfee-6ad9-...
Buddy
Column Name
Column Value
fead97e9-4d77-40c9-ba15-c45478542e20
owner
pet_name
642adfee-6ad9-...
Max
47045afb-fd11-44c6-9d0f-82428434e887
owner
pet_name
642adfee-6ad9-...
Rocky

Choosing a Partition Key

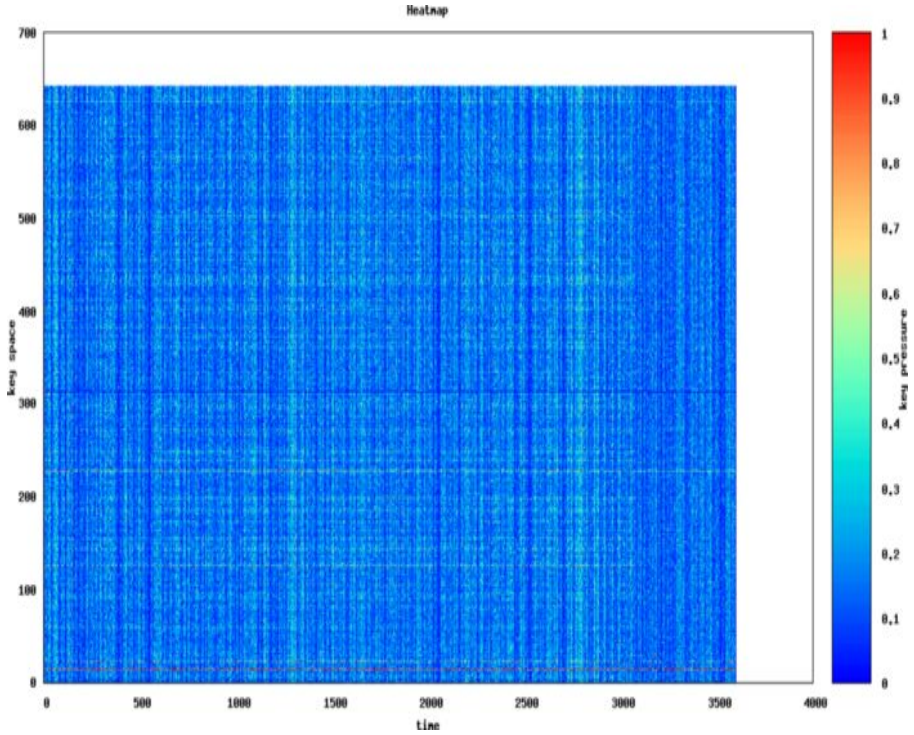
- User Name
- User ID
- User ID + Time
- Sensor ID
- Sensor ID + Time
- Customer

- State
- Age
- Favorite NBA Team
- Team Angel or Team Spike

https://commons.wikimedia.org/

Hot (top) Partition

[Nodetool top partitions]



Why Amazon DynamoDB isn't for everyone, Forrest Brazeal
https://acloudguru.com/blog/engineering/why-amazon-dynamodb-isnt-for-everyone-and-how-to-decide-when-it-s-for-you

Wide Partition Example

Query:

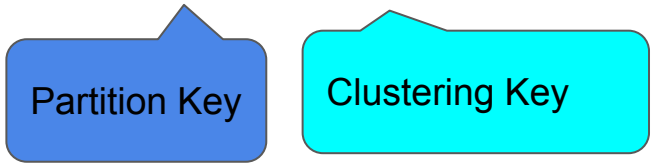
```
SELECT * from heartrate_v10 WHERE
pet_chip_id = 80d39c78-9dc0-11eb-a8b3-0242ac130003 LIMIT 1;
```

```
SELECT * from heartrate_v10 WHERE
pet_chip_id = 80d39c78-9dc0-11eb-a8b3-0242ac130003 AND
time >= '2021-05-01 01:00+0000'  AND
time <  '2021-05-01 01:03+0000';
```

https://gist.github.com/tzach/7486f1a0cc904c52f4514f20f14d2a97

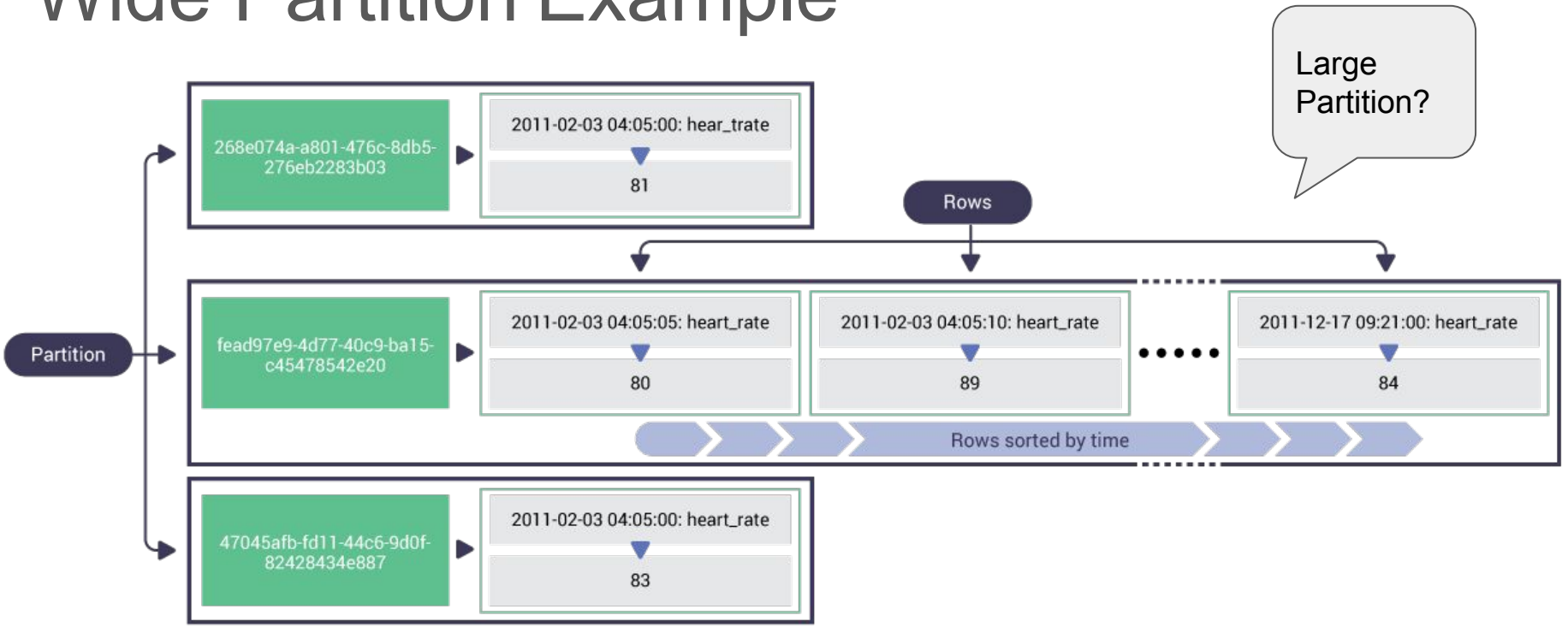
Wide Partition Example

```
CREATE TABLE heartrate_v10 (
    pet_chip_id uuid,
    owner uuid,
    time timestamp,
    heart_rate int,
    PRIMARY KEY (pet_chip_id, time)
);
```



pet_chip_id	time	heart_rate
80d39c78-9dc0-11eb-a8b3-0242ac130003	2021-05-01 01:00:00.000000+0000	120
80d39c78-9dc0-11eb-a8b3-0242ac130003	2021-05-01 01:01:00.000000+0000	121
80d39c78-9dc0-11eb-a8b3-0242ac130003	2021-05-01 01:02:00.000000+0000	120

Wide Partition Example



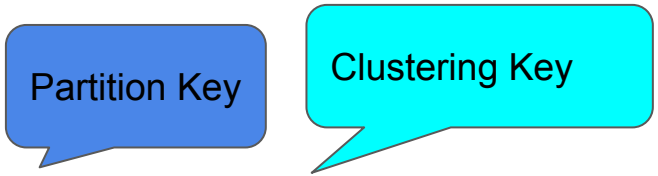
Choosing a Clustering Key

- Allow useful range queries
- Allow useful LIMIT

https://commons.wikimedia.org/

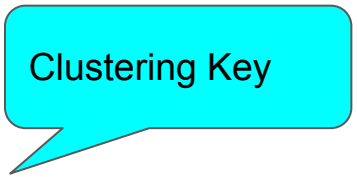
```
SELECT * from heartrate_v10 WHERE
pet_chip_id = 80d39c78-9dc0-11eb-a8b3-0242ac130003 LIMIT 1;
```

```
CREATE TABLE heartrate_v5 (
   pet_chip_id uuid,
   time timestamp,
   heart_rate int,
   PRIMARY KEY (pet_chip_id, time)
   ) WITH CLUSTERING ORDER BY (time DESC);
```



Too Wide Partition ?

```
CREATE TABLE heartrate_v6 (
   pet_chip_id uuid,
   date text,
   time timestamp,
   heart_rate int,
   PRIMARY KEY  ((pet_chip_id, date), time));
```



3.
Materialized Views and Secondary Index

Example

```
CREATE TABLE heartrate_v10 (
    pet_chip_id uuid,
    owner uuid,
    time timestamp,
    heart_rate int,
    PRIMARY KEY (pet_chip_id, time)
);
```

Partition Key

Clustering Key

https://gist.github.com/tzach/7486f1a0cc904c52f4514f20f14d2a97

Example

```
SELECT * FROM heartrate_v10 WHERE pet_chip_id =
a2a60505-3e17-4ad4-8e1a-f11139caa1cc;
```

```
SELECT * FROM heartrate_v10 WHERE owner =
642adfee-6ad9-4ca5-aa32-a72e506b8ad8;
```

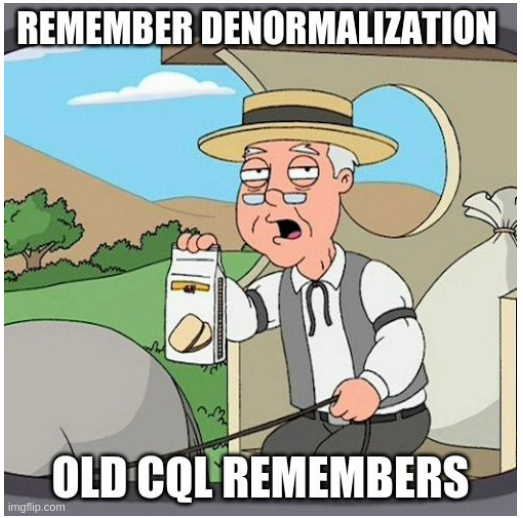
```
SELECT * FROM heartrate_v10 WHERE owner =
642adfee-6ad9-4ca5-aa32-a72e506b8ad8  ALLOW FILTERING;
```

https://gist.github.com/tzach/4b9dadbc6e8a9c50369da05631c5e13e

CQL syntax

```
CREATE MATERIALIZED VIEW heartrate_by_owner AS
 SELECT * FROM heartrate_v10
 WHERE owner IS NOT NULL AND pet_chip_id IS NOT NULL AND time IS NOT NULL
 PRIMARY KEY(owner, pet_chip_id, time);
```

```
SELECT * FROM heartrate_by_owner WHERE owner =
642adfee-6ad9-4ca5-aa32-a72e506b8ad8;
```



https://docs.scylladb.com/getting-started/mv/

Example

Base Table

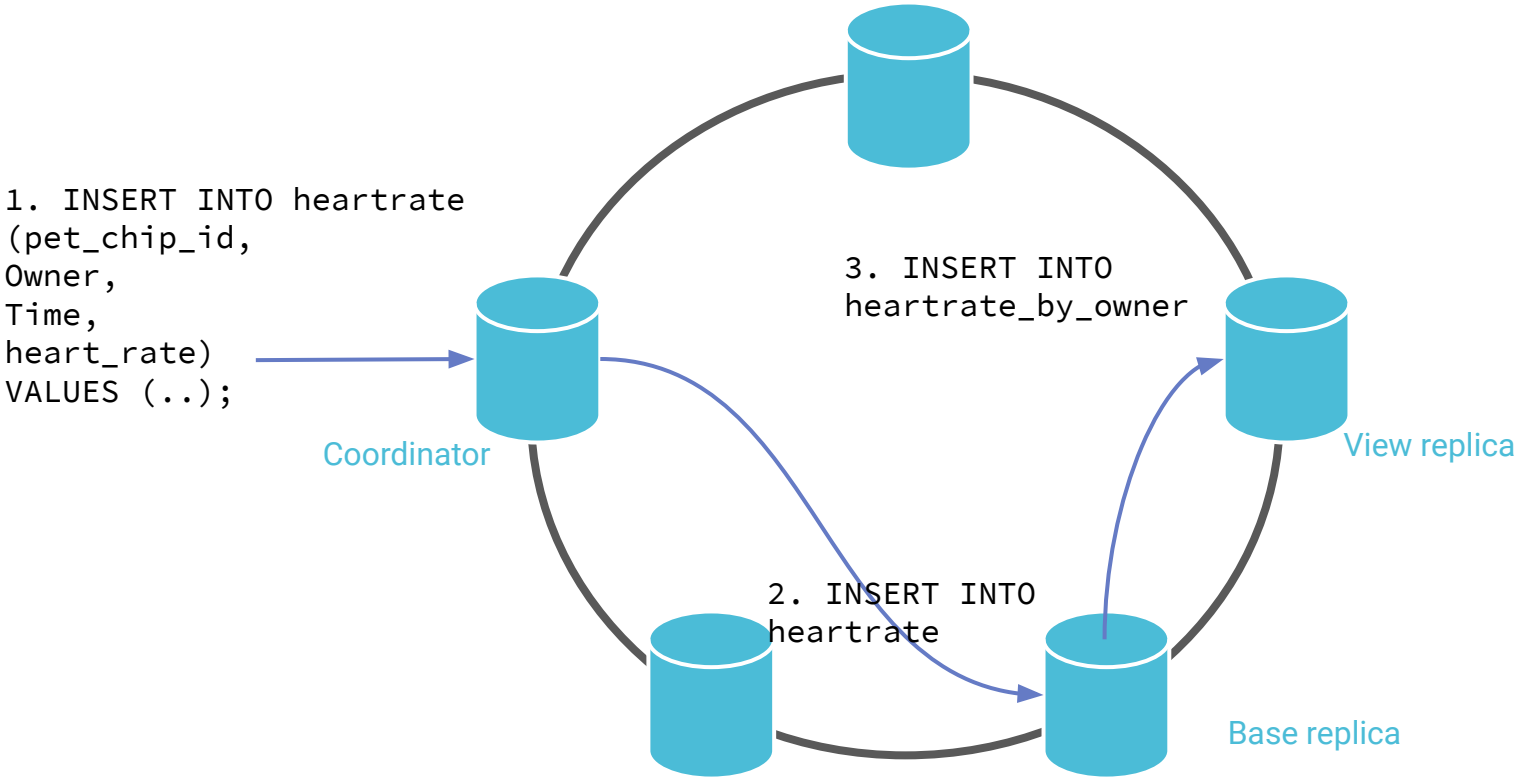
pet_chip_id	time	owner	heart_rate
80d39c78-9dc0-11eb-a8b3-0242ac130003	2021-05-01 01:00:00.000000+0000	642adfee-6ad9-4ca5-aa32-a72e506b8ad8	120
80d39c78-9dc0-11eb-a8b3-0242ac130003	2021-05-01 01:01:00.000000+0000	642adfee-6ad9-4ca5-aa32-a72e506b8ad8	121
80d39c78-9dc0-11eb-a8b3-0242ac130003	2021-05-01 01:02:00.000000+0000	642adfee-6ad9-4ca5-aa32-a72e506b8ad8	12

CREATE MATERIALIZED VIEW

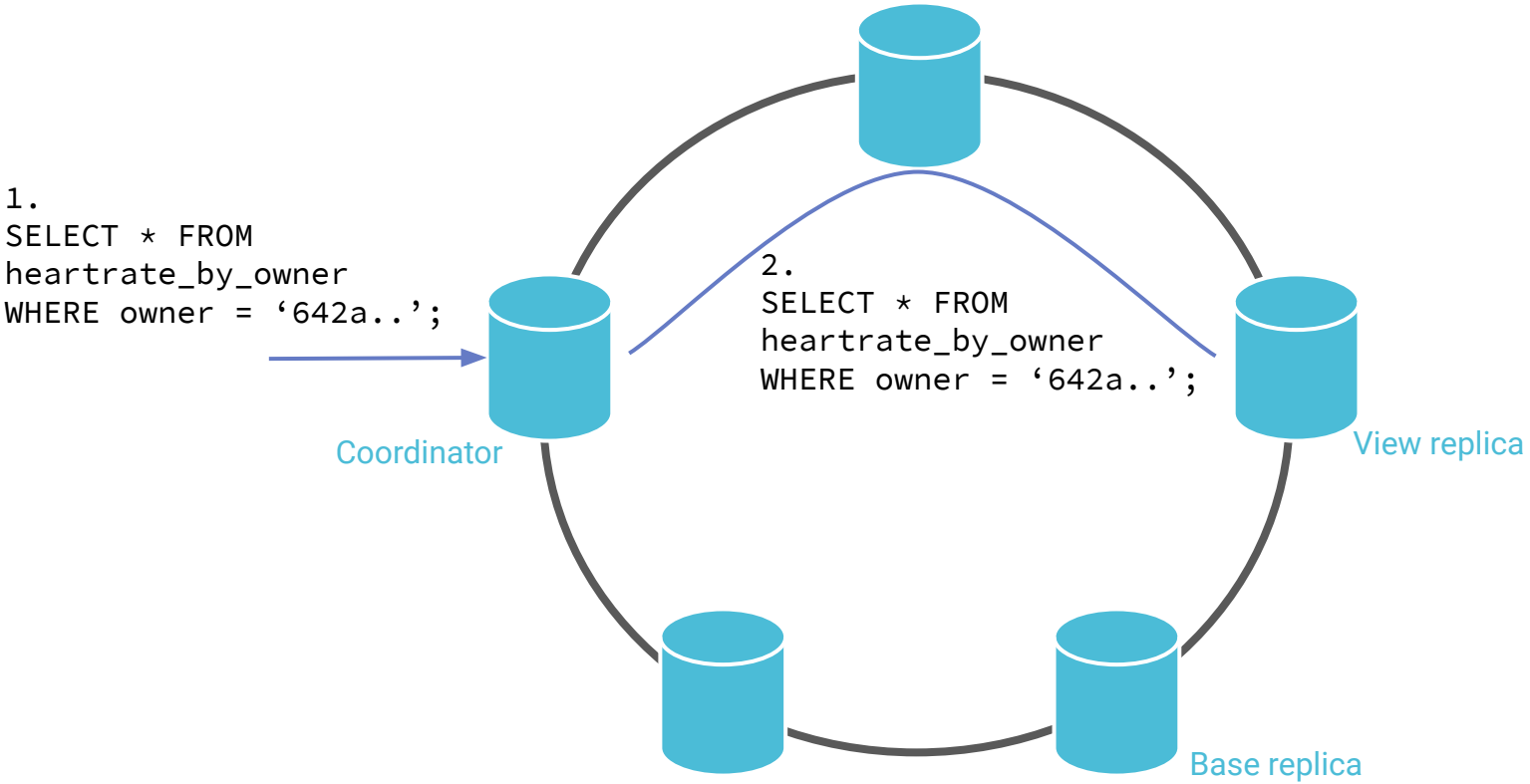
View

owner	pet_chip_id	time	heart_rate
642adfee-6ad9-4ca5-aa32-a72e506b8ad8	80d39c78-9dc0-11eb-a8b3-0242ac130003	2021-05-01 01:00:00.000000+0000	120
642adfee-6ad9-4ca5-aa32-a72e506b8ad8	80d39c78-9dc0-11eb-a8b3-0242ac130003	2021-05-01 01:01:00.000000+0000	121
642adfee-6ad9-4ca5-aa32-a72e506b8ad8	80d39c78-9dc0-11eb-a8b3-0242ac130003	2021-05-01 01:02:00.000000+0000	120

View is another table



View is another table

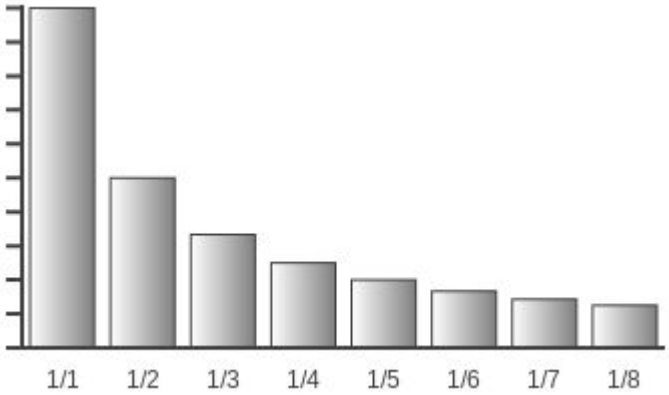


Choosing a MV Index Key

- High Cardinality
- Even Distribution

Avoid

- Low Cardinality
- Hot Partition
- Large Partition



https://www.codedrome.com/zipfs-law-in-python/

Wait, what about JOINs?

- Try to avoid them using denormalization
- For ad-hoc operations - use external engine: Spark, Presto

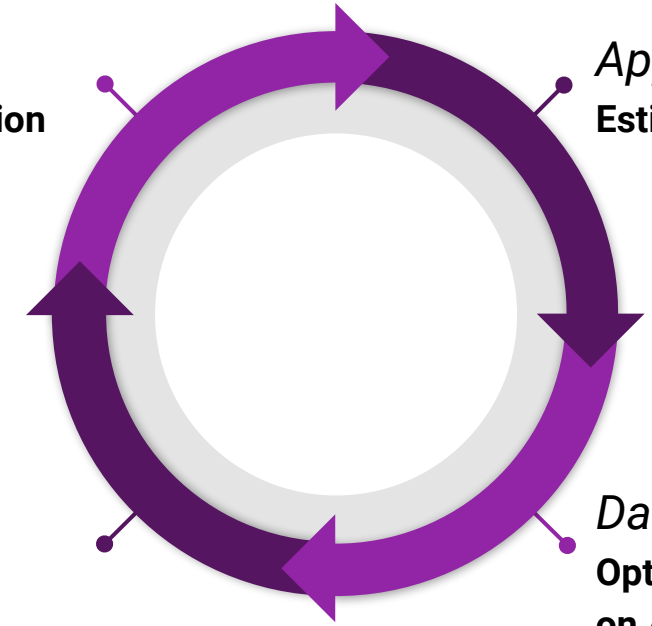
Circle of (NoSQL) Life

Update

Update schema / application

Application

Estimate Read and Write pattern



Measure

Use Metric to evaluate the data model

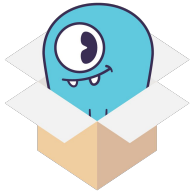
Data Model

Optimize the model base on Apps SLA

Experience Scylla for Yourself

The Speaker's camera displays here

Download Scylla Open Source:
scylladb.com/download



Learn more https://university.scylladb.com/
Contact me tzach@scylladb.com

THANK YOU !

PERCONA
LIVEONLINE
MAY 12 - 13th
2021