AtlasDB

DANIELLE KRAMER, Engineering Lead <dkramer@palantir.com>

ARI GESHER, Engineering Ambassador <agesher@palantir.com>
@alephbass

Palantir

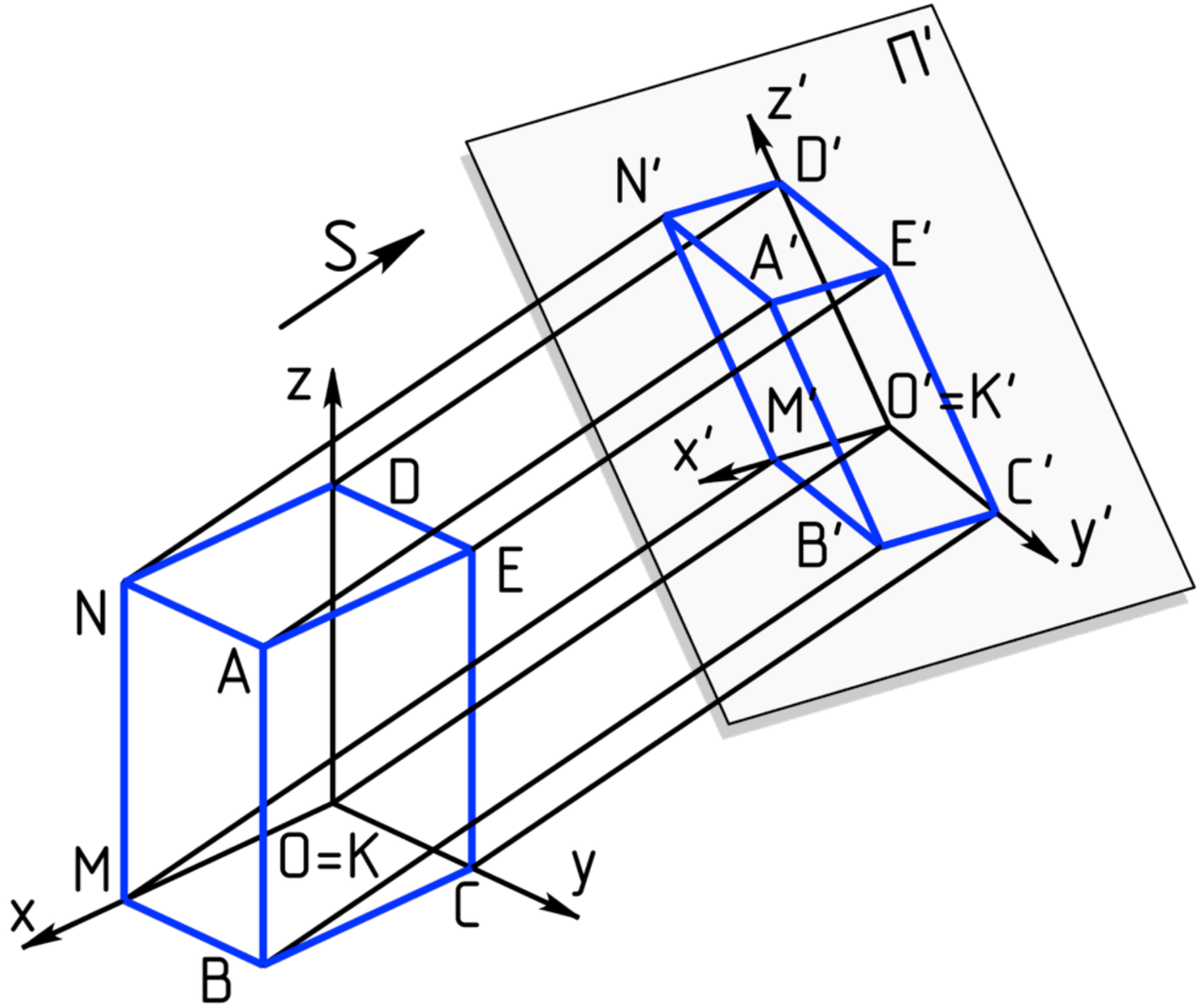
OVERVIEW

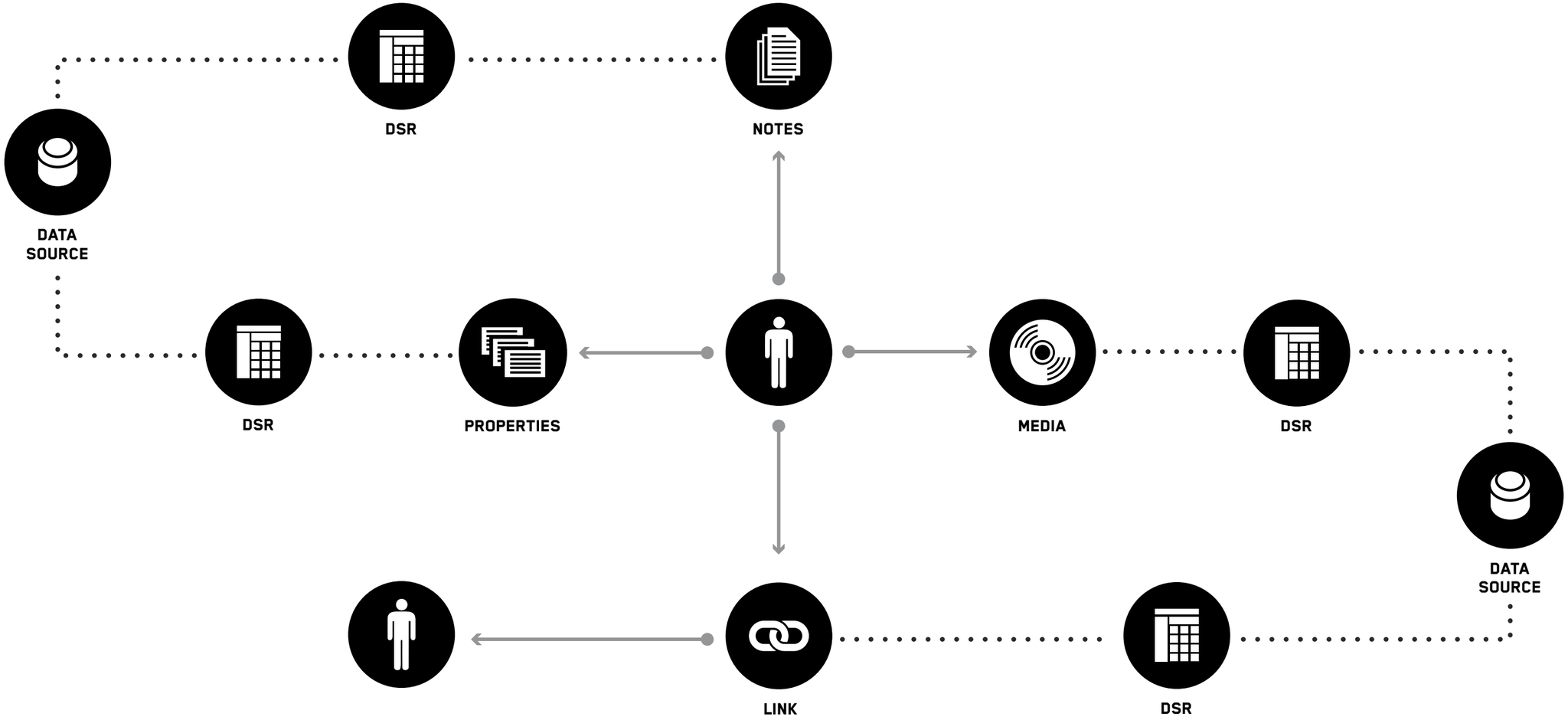
- Data at Palantir
- Why did we build AtlasDB?
- What is AtlasDB?
- How does AtlasDB work?
- Results & Benchmarks
- What's next?

OVERVIEW

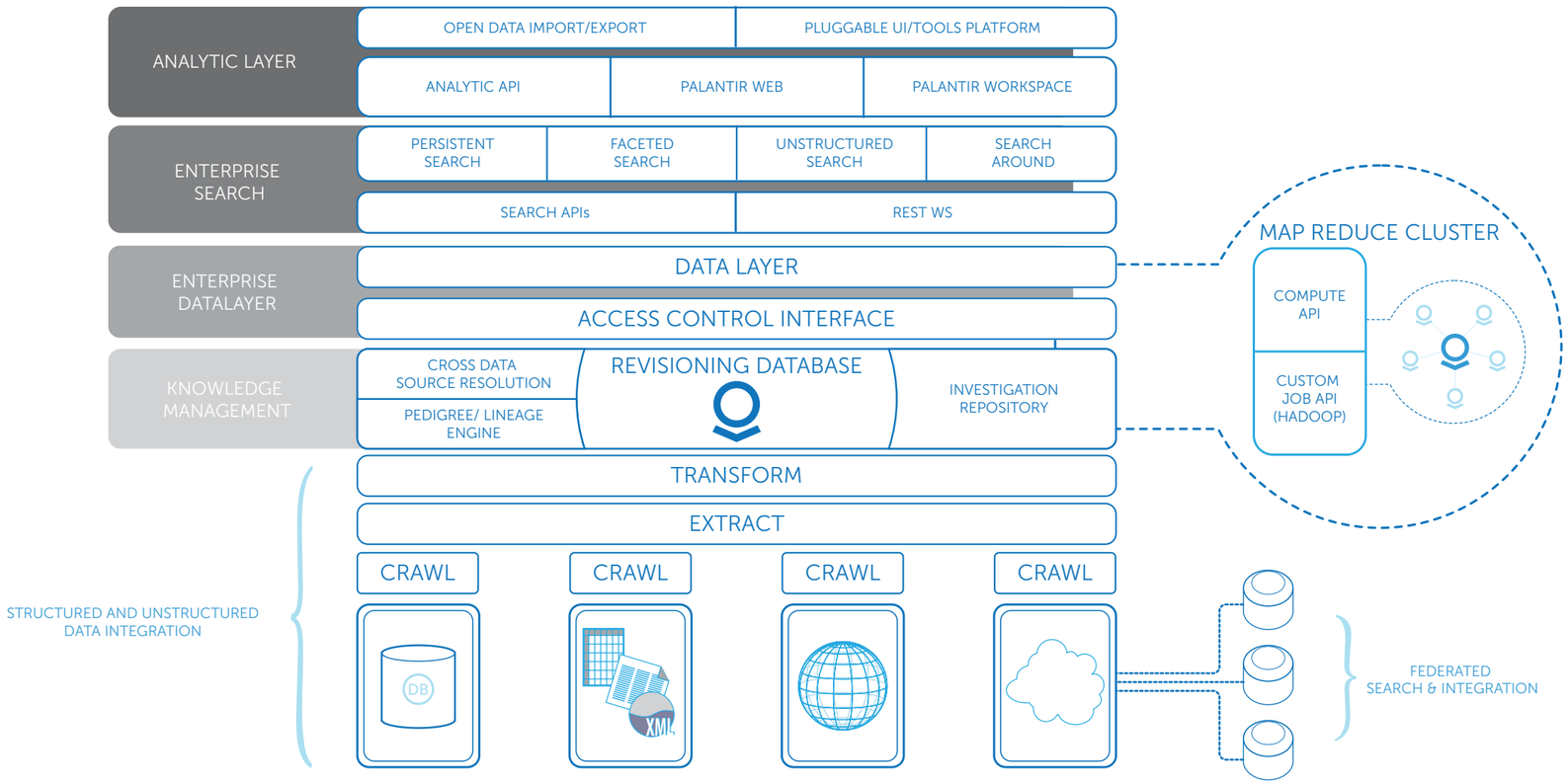
- **Data at Palantir**
- Why did we build AtlasDB?
- What is AtlasDB?
- How does AtlasDB work?
- Results & Benchmarks
- What's next?

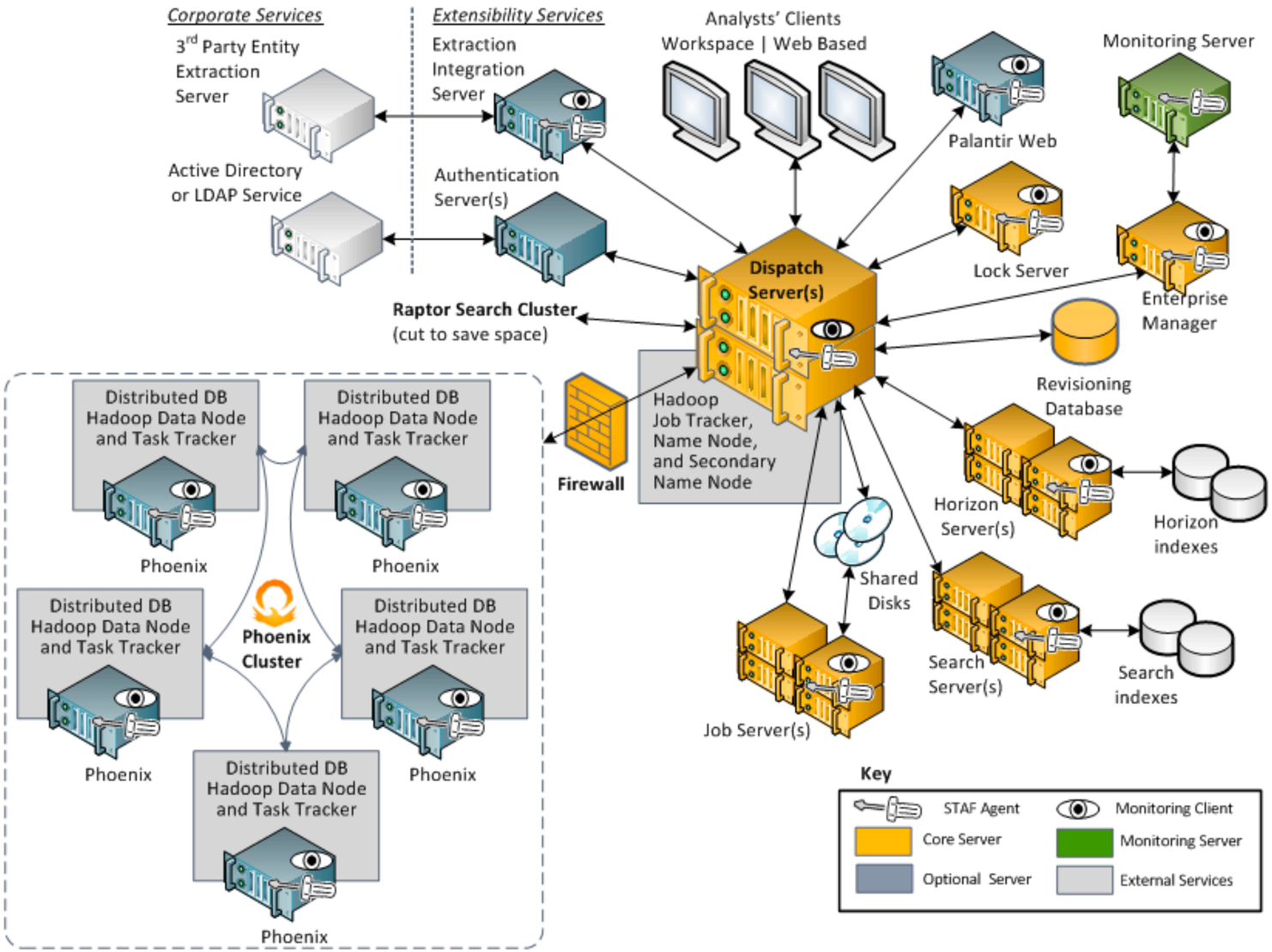

Palantir


П′
z′
N′
D′
A′
E′
S
z
M′
O′=K′
x′
C′
y′
B′
D
E
N
A
M
O=K
x
y
C
B


DSR
NOTES
DATA SOURCE
DSR
PROPERTIES
MEDIA
DSR
DATA SOURCE
LINK
DSR

INTELLIGENCE INFRASTRUCTURE

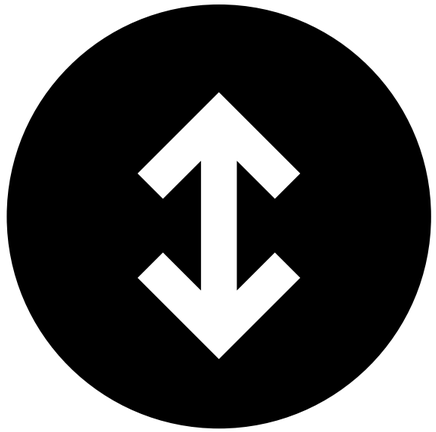



Corporate Services
3rd Party Entity Extraction Server
Active Directory or LDAP Service
Extensibility Services
Extraction Integration Server
Authentication Server(s)
Analysts' Clients
Workspace | Web Based
Palantir Web
Monitoring Server
Lock Server
Enterprise Manager
Dispatch Server(s)
Raptor Search Cluster
(cut to save space)
Revisioning Database
Firewall
Hadoop Job Tracker, Name Node, and Secondary Name Node
Horizon Server(s)
Horizon indexes
Shared Disks
Search Server(s)
Search indexes
Job Server(s)
Distributed DB Hadoop Data Node and Task Tracker
Phoenix
Phoenix Cluster
Key
STAF Agent
Monitoring Client
Core Server
Monitoring Server
Optional Server
External Services

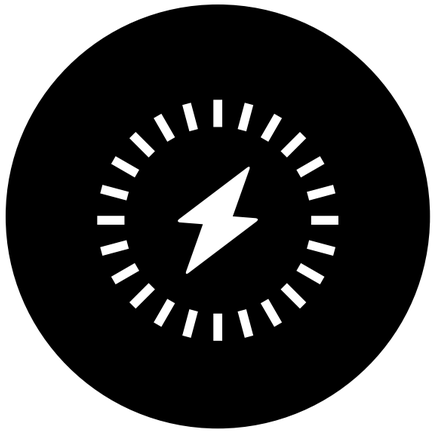
OVERVIEW

- Data at Palantir
- **Why did we build AtlasDB?**
- What is AtlasDB?
- How does AtlasDB work?
- Results & Benchmarks
- What's next?

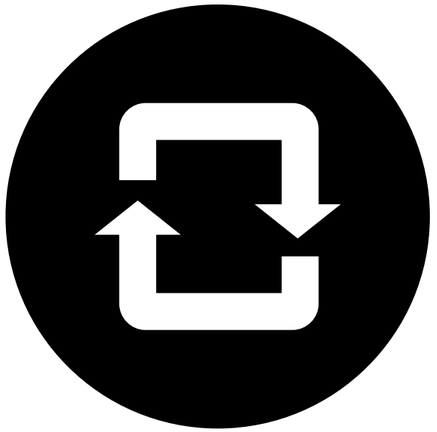
WHAT DO WE CARE ABOUT IN OUR DATA STORE?



SCALE



SPEED



TRANSACTIONS

LIMITATIONS IN CURRENT SOLUTIONS

	ORACLE	NOSQL STORES
SCALE	N	Y
SPEED	Y	Y
TRANSACTIONS	Y	N

OVERVIEW

- Data at Palantir
- Why did we build AtlasDB?
- **What is AtlasDB?**
- How does AtlasDB work?
- Results & Benchmarks
- What's next?

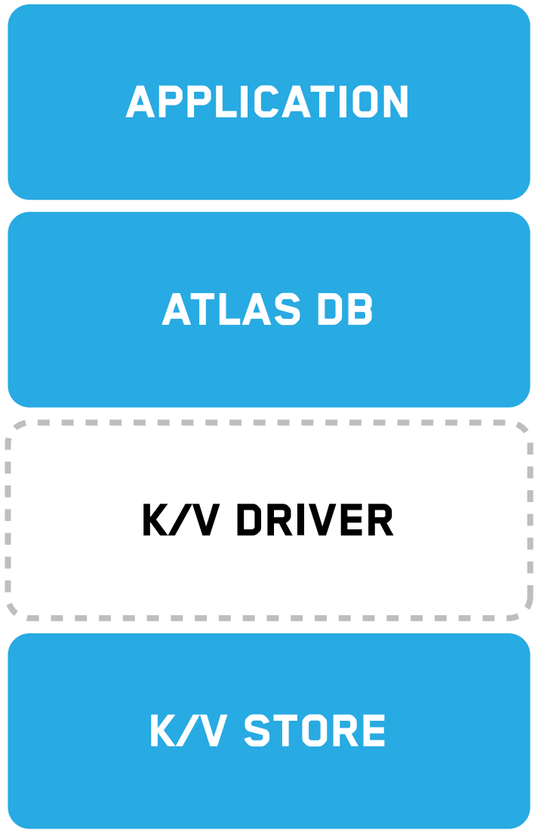
WHAT IS ATLASDB?

AtlasDB is a ***transactional API for key-value stores***.

TRANSACTIONS

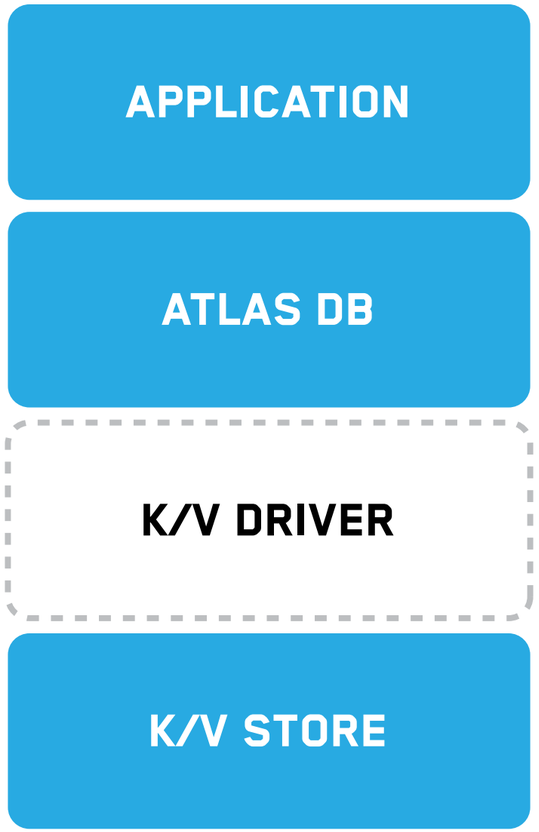
- Atomic
- Consistent
- Isolated
- Durable

ATLAS DB ARCHITECTURE



ATLAS DB ARCHITECTURE

- Don't have to build a database
- Easy to change database
- Different DBs for different scale



API DETAILS

- Java
- Column Store
- Strongly Typed

OVERVIEW

- Data at Palantir
- Why did we build AtlasDB?
- What is AtlasDB?
- **How does AtlasDB work?**
- Results & Benchmarks
- What's next?

PERCOLATOR

- *Large-scale Incremental Processing Using Distributed Transactions and Notifications*, Peng and Dabek 2010
- ACID transactions on top of Bigtable
- Thousands of machines
- Writes are 75% slower
- Only at Google

ATLAS DB WRITE PROTOCOL

- Buffer data in memory with a write timestamp

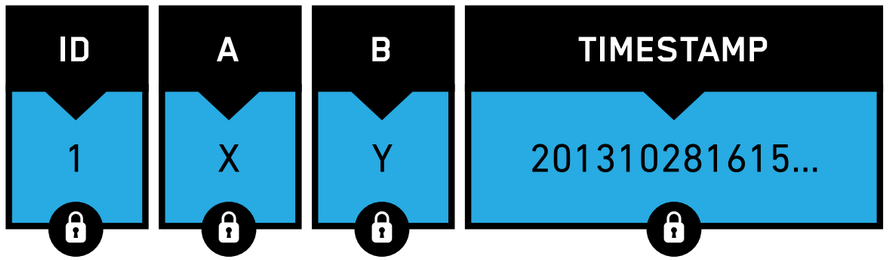
Data Table: In Memory

ID	A	B	TIMESTAMP
1	X	Y	201310281615...

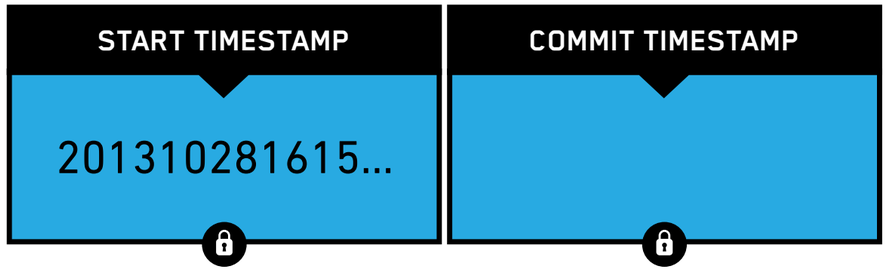
ATLAS DB WRITE PROTOCOL

- Lock data rows and transaction row

Data Table: In Memory



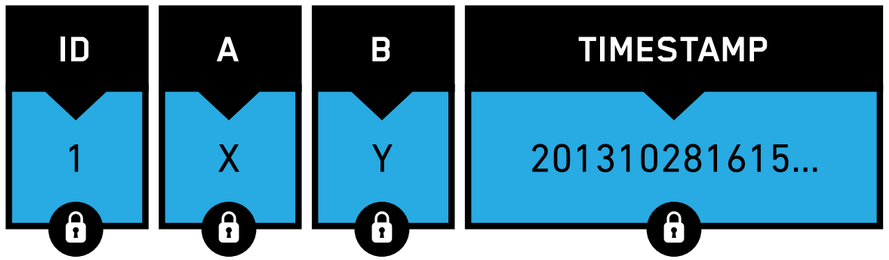
Transaction Table : In Memory



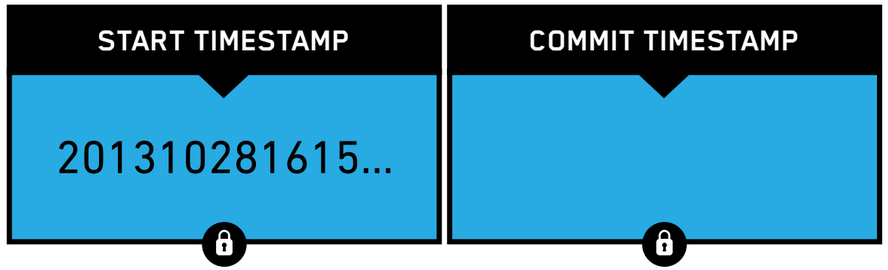
ATLAS DB WRITE PROTOCOL

- Write data rows

Data Table: On Disk



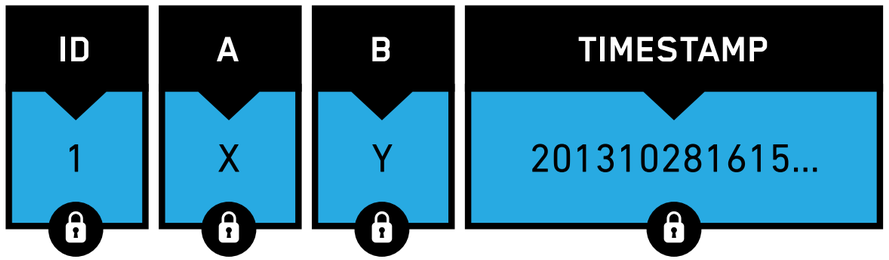
Transaction Table : In Memory



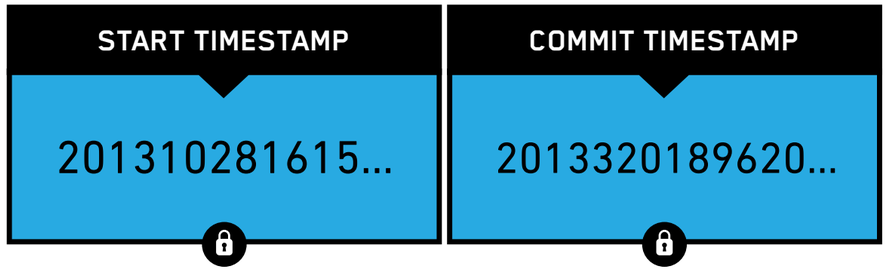
ATLAS DB WRITE PROTOCOL

- Write transaction row with commit timestamp

Data Table: On Disk



Transaction Table: On Disk



ATLAS DB WRITE PROTOCOL

- Release lock

Data Table: On Disk

ID	A	B	TIMESTAMP
1	X	Y	201310281615...

Transaction Table: On Disk

START TIMESTAMP	COMMIT TIMESTAMP
201310281615...	2013320189620...

ATLAS DB WRITE PROTOCOL

- Buffer data in memory with a write timestamp
- Lock data rows and transaction row
- Write data rows
- Write transaction row with commit timestamp
- Release locks

ATLAS DB COMPONENTS

- Lock Server
- Timestamp Server
- Transaction Server
- Key-Value Store

KEY POINTS

- Lock Server is in memory
  - Reduces write amplification
- Transaction metadata is lightweight
  - Low write overhead
- Separate transaction table
  - Simplifies commit failure cleanup

OVERVIEW

- Data at Palantir
- Why did we build AtlasDB?
- What is AtlasDB?
- How does AtlasDB work?
- **Results & Benchmarks**
- What's next?

THE BIG RESULT

- Transactions with 15% Overhead

SPEED BENCHMARKS IN PRACTICE

- Object loads - 250% faster
- Object stores - 15% faster

SCALE RESULTS

- Performance on Cassandra was constant from 1 to 10 nodes

STATUS

- Dark Launch - 6 months ago
- Oracle - 3 months ago
- Postgres - last month
- Cassandra - next month

RESULT SUMMARY

- 3x as much data today
- 10-100x more data in the future
- 50%-250% improved performance compared to Oracle

OVERVIEW

- Data at Palantir
- Why did we build AtlasDB?
- What is AtlasDB?
- How does AtlasDB work?
- Results & Benchmarks
- **What's next?**

WHAT'S NEXT?

- Continued scaling
- Performance tuning
- Wider usage

OPEN SOURCING ATLAS DB

- Targeting 2014 release
- Simplifying and improving API
- Documenting and providing examples

THANK YOU