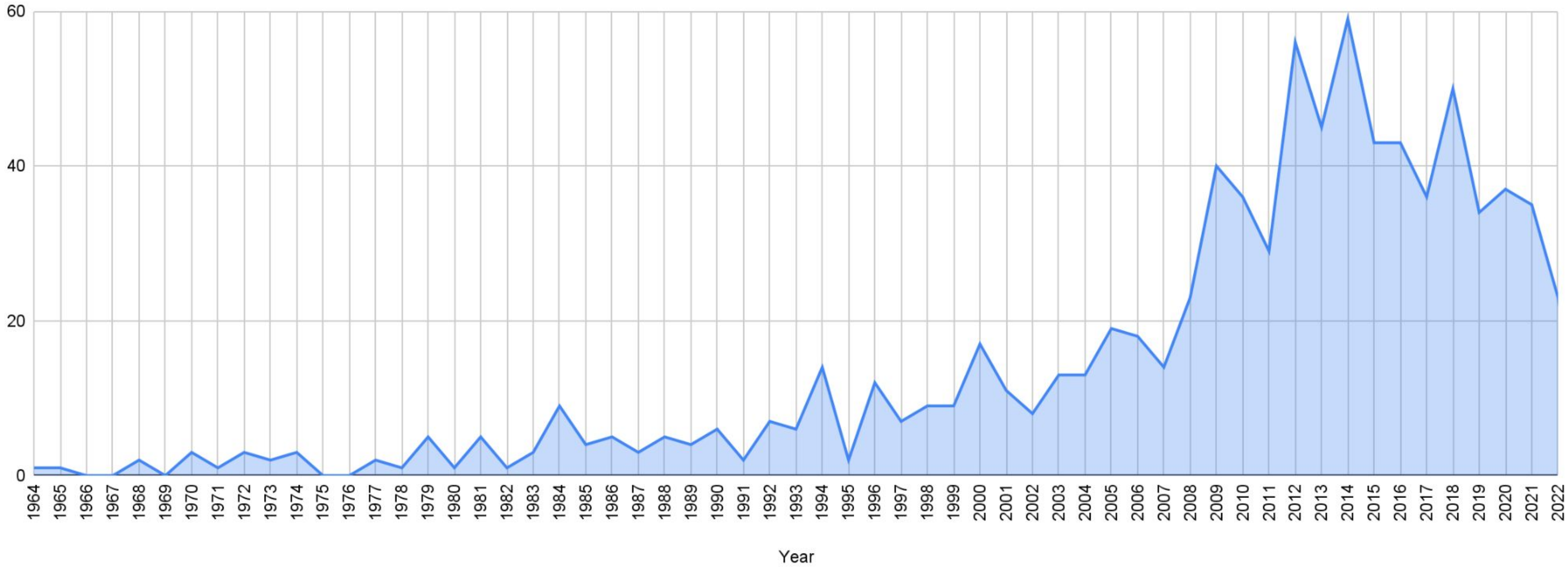
Unbundling of the DBMS stack

Mosha Pasumansky & Benjamin Wagner

DBMSs created per Year



Source: dbdb.io

CeresDB	Time series database	Query Engine - DataFusion WAL - RocksDB, OceanBase, Memtable - AgateDB SST - derived from Parquet
LingoDB	Data processing system that leverages compiler technology	Parser - libpg_query
CnosDB	Time series database	RPC - ArrowFlight, Query Engine - DataFusion
RisingWave	Distributed SQL for stream processing	DataFusion
MonographDB	Multi model database	Compute - MariaDB, Storage - Cassandra
nucliadb	AI search / generative answers / vector database	LMBD and/or TiKV
spicedb	Database for managing security permissions checking	CockroachDB
Dragonfly	Redis replacement	
Oriole data base	Next gen storage engine for PostgreSQL	PostgreSQL extension
EdgelessDB	Database for confidential computing (inside SGX enclave)	Forked MariaDB, Storage engine - RocksDB
NEON	Serverless PostgreSQL	PostgreSQL
FerretDB	MongoDB alternative	PostgreSQL

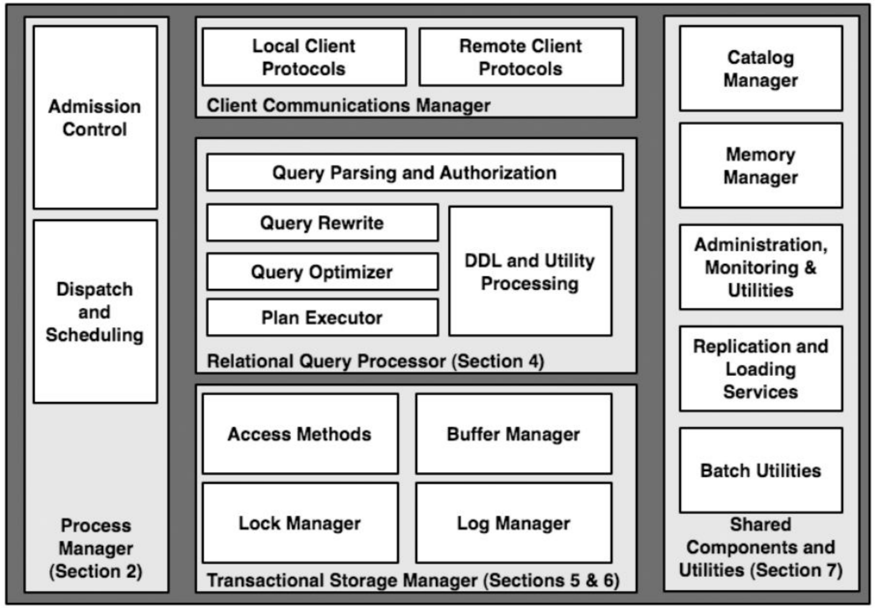
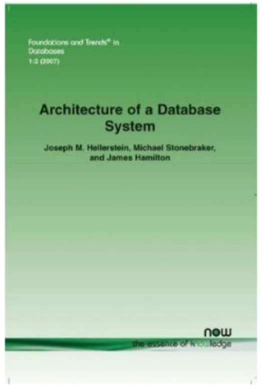

Admission Control
Dispatch and Scheduling
Process Manager (Section 2)
Local Client Protocols
Remote Client Protocols
Client Communications Manager
Query Parsing and Authorization
Query Rewrite
Query Optimizer
Plan Executor
DDL and Utility Processing
Relational Query Processor (Section 4)
Access Methods
Buffer Manager
Lock Manager
Log Manager
Transactional Storage Manager (Sections 5 & 6)
Catalog Manager
Memory Manager
Administration, Monitoring & Utilities
Replication and Loading Services
Batch Utilities
Shared Components and Utilities (Section 7)

Fig. 1.1 Main components of a DBMS.



Foundations and Trends® in
Databases
Vol. 1, No. 2 (2007) 141–259

DOI: 10.1561/1900000002

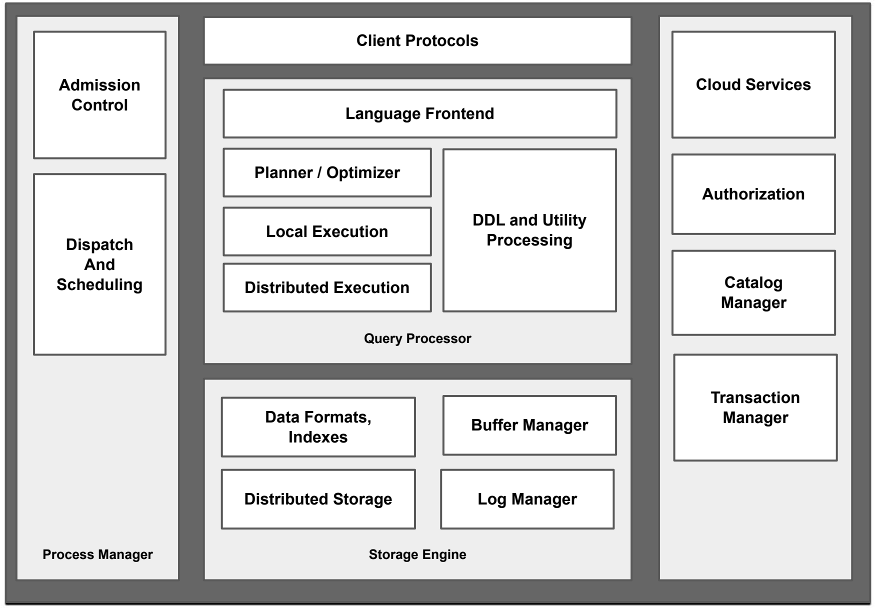
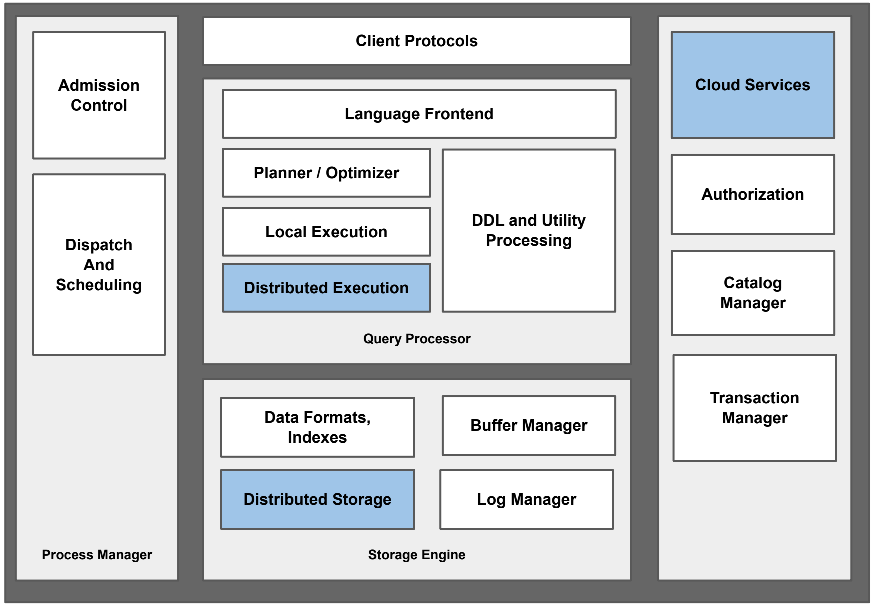
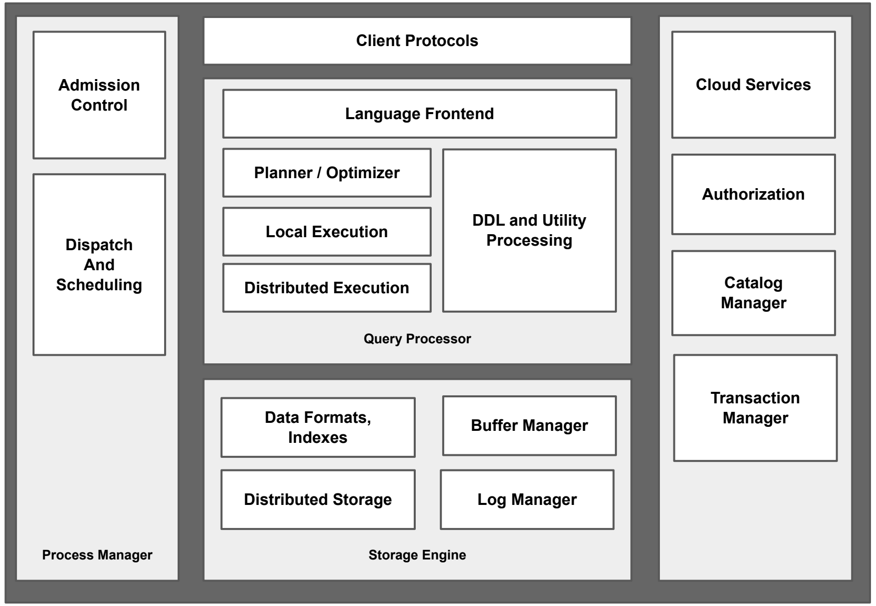
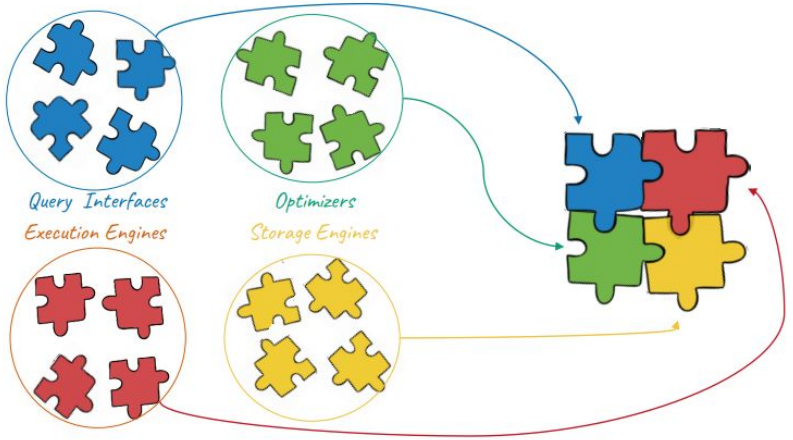


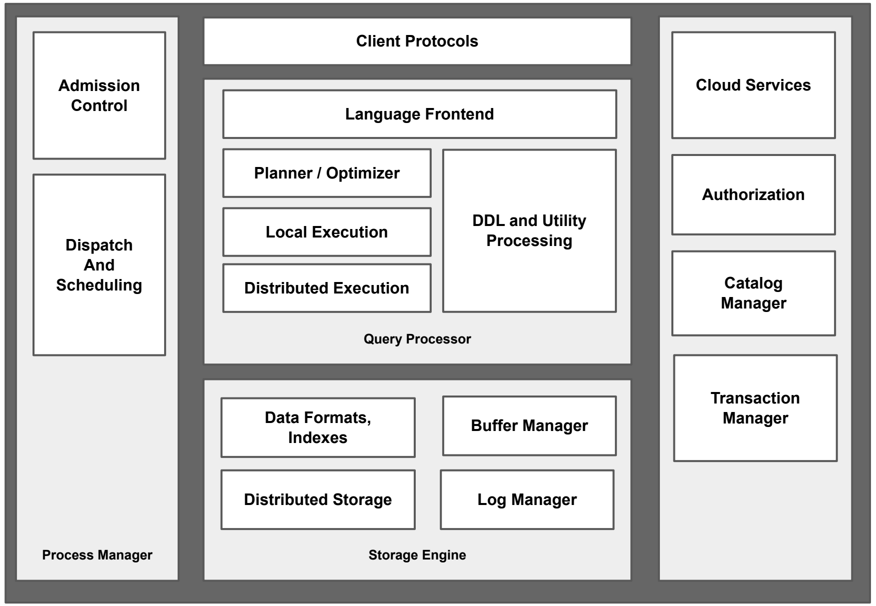
Fig. 1.1 Main components of a DBMS.

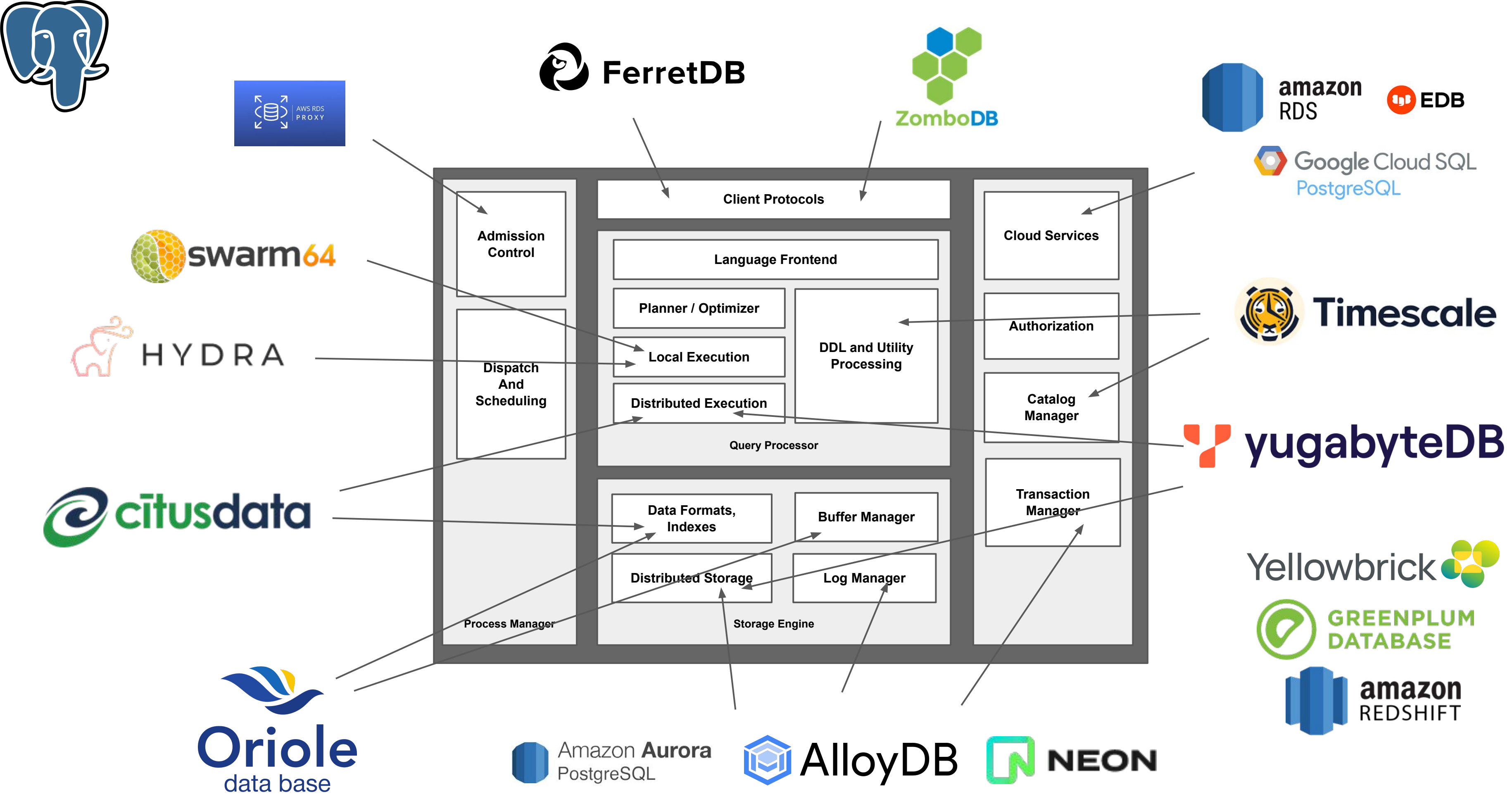

Admission Control
Dispatch And Scheduling
Process Manager
Client Protocols
Language Frontend
Planner / Optimizer
Local Execution
Distributed Execution
DDL and Utility Processing
Query Processor
Data Formats, Indexes
Buffer Manager
Distributed Storage
Log Manager
Storage Engine
Cloud Services
Authorization
Catalog Manager
Transaction Manager

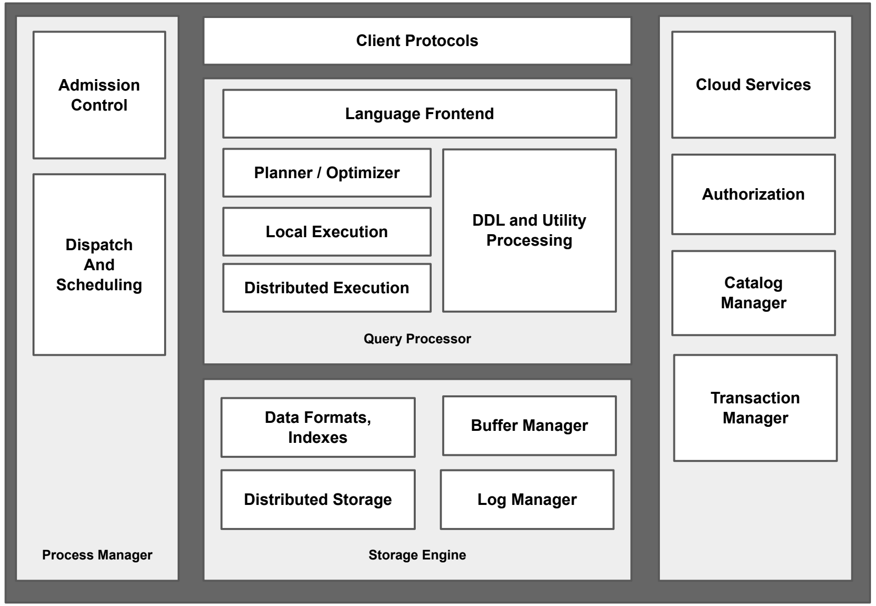

Admission Control
Dispatch And Scheduling
Process Manager
Client Protocols
Language Frontend
Planner / Optimizer
Local Execution
Distributed Execution
DDL and Utility Processing
Query Processor
Data Formats, Indexes
Buffer Manager
Distributed Storage
Log Manager
Storage Engine
Cloud Services
Authorization
Catalog Manager
Transaction Manager

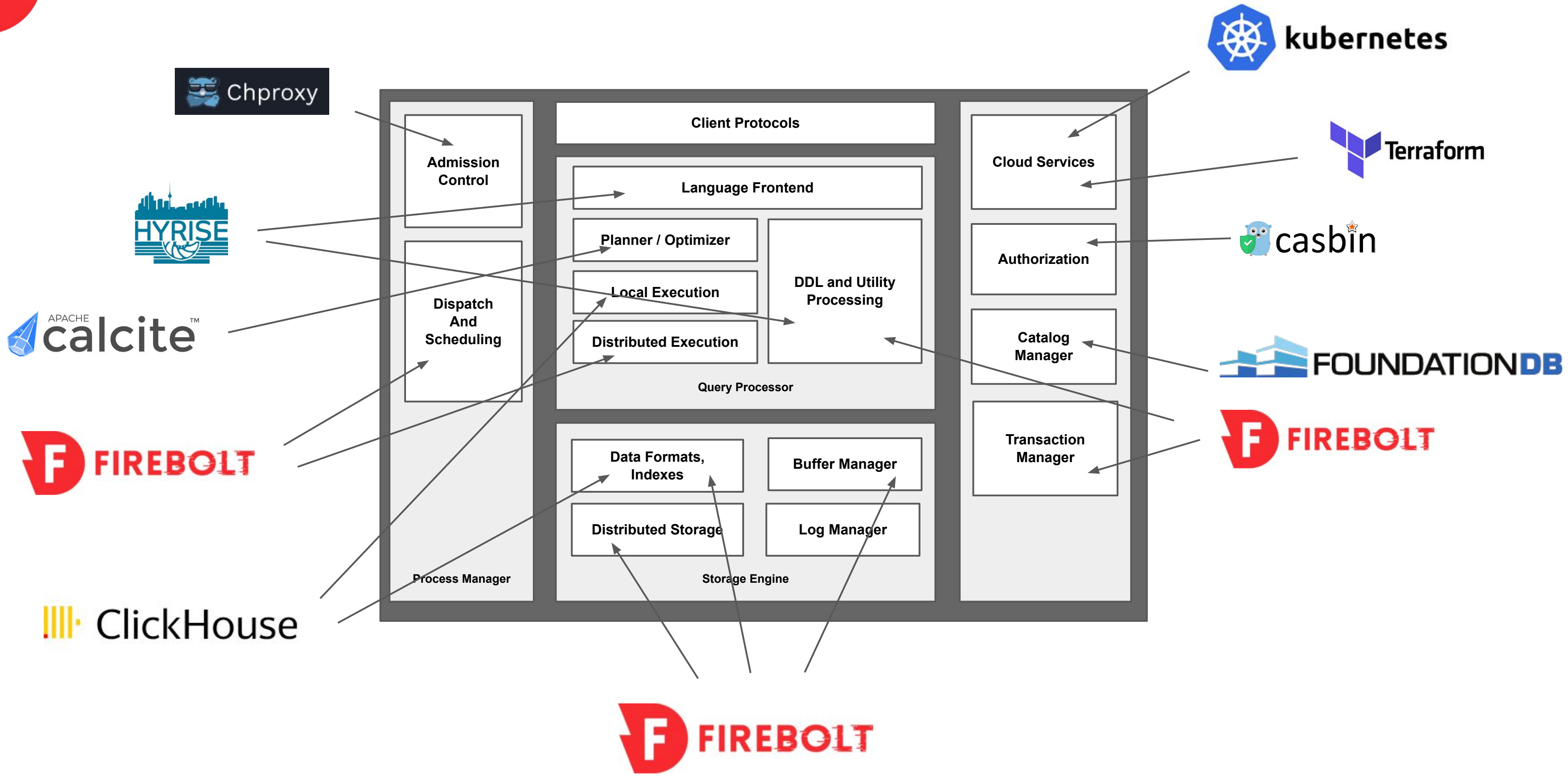


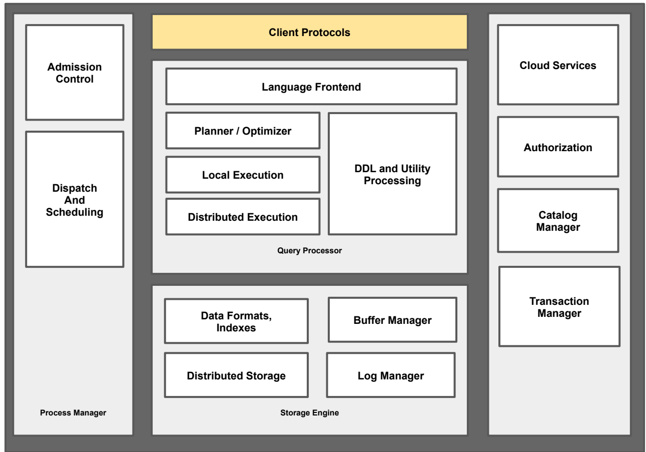
Credit: Towards a Modular Data Management System Framework
https://cdmsworkshop.github.io/2022/Slides/Fri_C2.5_HaralamposGavriilidis.pdf

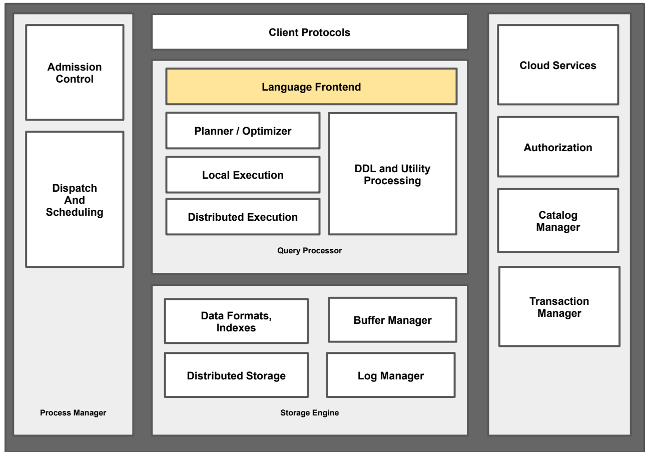

Admission Control
Dispatch And Scheduling
Process Manager
Client Protocols
Language Frontend
Planner / Optimizer
Local Execution
Distributed Execution
DDL and Utility Processing
Query Processor
Data Formats, Indexes
Buffer Manager
Distributed Storage
Log Manager
Storage Engine
Cloud Services
Authorization
Catalog Manager
Transaction Manager


AWS RDS PROXY
FerretDB
ZomboDB
amazon RDS
EDB
Google Cloud SQL PostgreSQL
swarm64
HYDRA
Timescale
yugabyteDB
citusdata
Yellowbrick
GREENPLUM DATABASE
amazon REDSHIFT
Oriole data base
Amazon Aurora PostgreSQL
AlloyDB
NEON
Admission Control
Dispatch And Scheduling
Process Manager
Client Protocols
Language Frontend
Planner / Optimizer
Local Execution
Distributed Execution
DDL and Utility Processing
Query Processor
Data Formats, Indexes
Buffer Manager
Distributed Storage
Log Manager
Storage Engine
Cloud Services
Authorization
Catalog Manager
Transaction Manager


Client Protocols
Admission Control
Dispatch And Scheduling
Process Manager
Language Frontend
Planner / Optimizer
Local Execution
Distributed Execution
DDL and Utility Processing
Query Processor
Data Formats, Indexes
Buffer Manager
Distributed Storage
Log Manager
Storage Engine
Cloud Services
Authorization
Catalog Manager
Transaction Manager

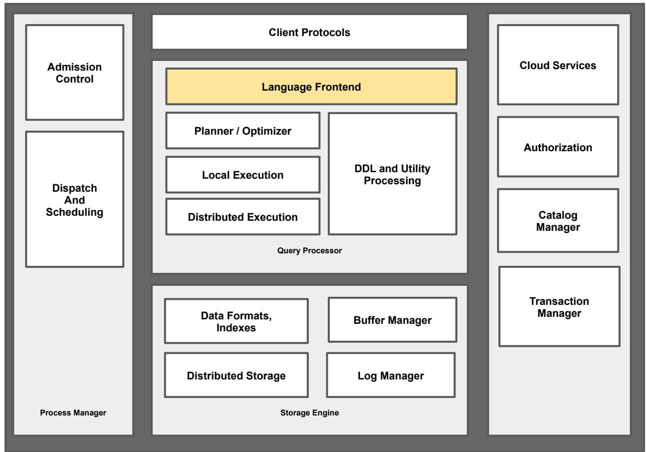

Chproxy
kubernetes
Terraform
casbin
HYRISE
APACHE calcite
FOUNDATIONDB
FIREBOLT
FIREBOLT
ClickHouse
FIREBOLT
Admission Control
Client Protocols
Cloud Services
Language Frontend
Planner / Optimizer
Authorization
Local Execution
DDL and Utility Processing
Dispatch And Scheduling
Distributed Execution
Catalog Manager
Query Processor
Transaction Manager
Data Formats, Indexes
Buffer Manager
Distributed Storage
Log Manager
Process Manager
Storage Engine


Admission Control
Dispatch And Scheduling
Process Manager
Client Protocols
Language Frontend
Planner / Optimizer
Local Execution
Distributed Execution
DDL and Utility Processing
Query Processor
Data Formats, Indexes
Buffer Manager
Distributed Storage
Log Manager
Storage Engine
Cloud Services
Authorization
Catalog Manager
Transaction Manager


Admission Control
Dispatch And Scheduling
Process Manager
Client Protocols
Language Frontend
Planner / Optimizer
Local Execution
Distributed Execution
DDL and Utility Processing
Query Processor
Data Formats, Indexes
Buffer Manager
Distributed Storage
Log Manager
Storage Engine
Cloud Services
Authorization
Catalog Manager
Transaction Manager

1. What language ?

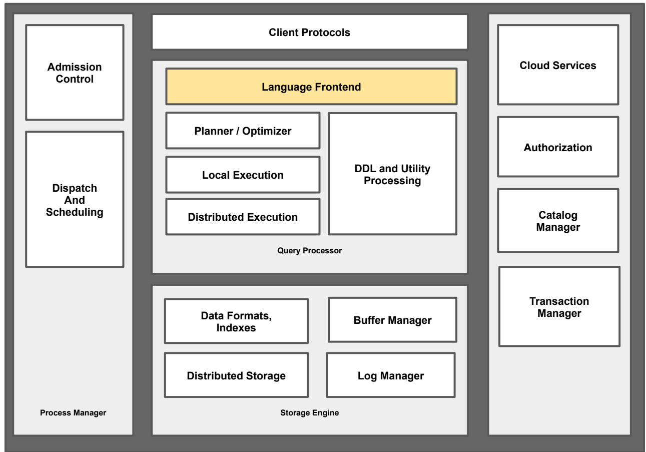
2. What language ?



1. What language ?

SQL PRQL PartiQL

2. What language ?



1. What language ?

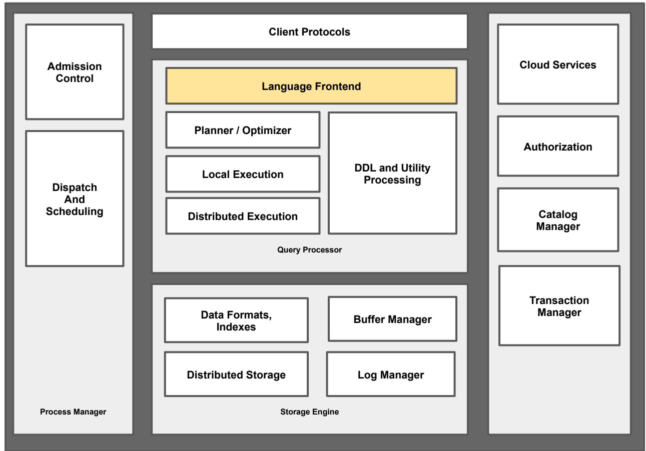
SQL PRQL PartiQL

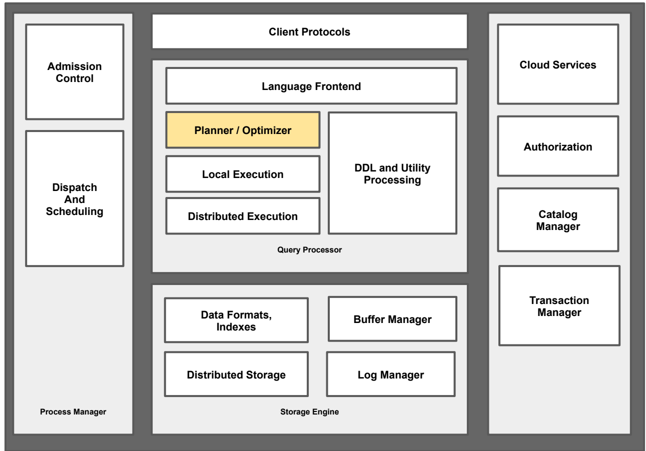
2. What language ?

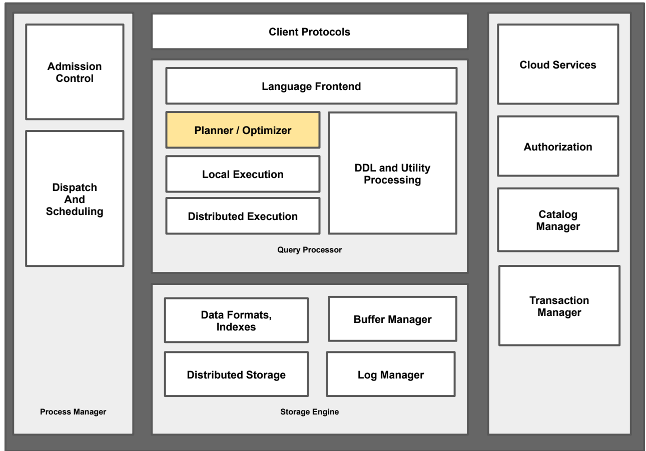
The Rust Programming Language

Java

GO




Admission Control
Dispatch And Scheduling
Process Manager
Client Protocols
Language Frontend
Planner / Optimizer
Local Execution
Distributed Execution
DDL and Utility Processing
Query Processor
Data Formats, Indexes
Buffer Manager
Distributed Storage
Log Manager
Storage Engine
Cloud Services
Authorization
Catalog Manager
Transaction Manager

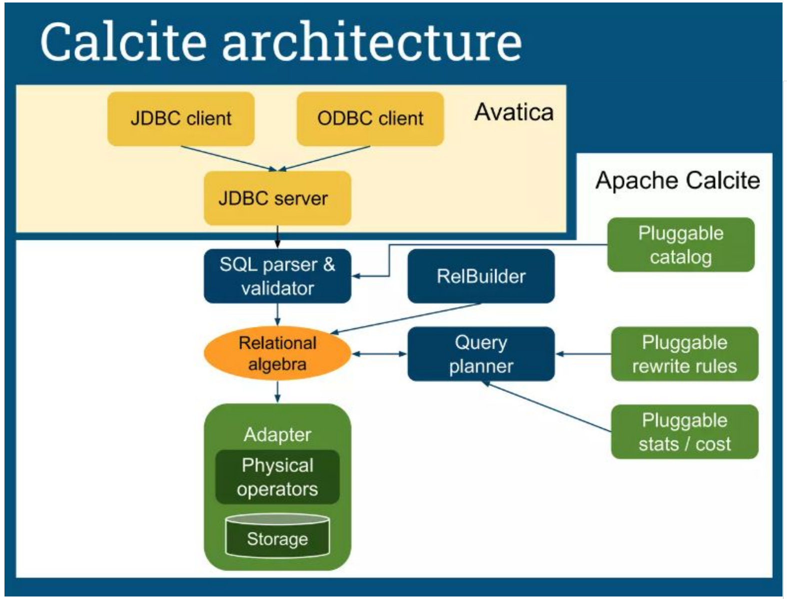

Client Protocols
Admission Control
Dispatch And Scheduling
Process Manager
Language Frontend
Planner / Optimizer
Local Execution
Distributed Execution
DDL and Utility Processing
Query Processor
Data Formats, Indexes
Buffer Manager
Distributed Storage
Log Manager
Storage Engine
Cloud Services
Authorization
Catalog Manager
Transaction Manager

APACHE calcite™

DATA FUSION

HYRISE

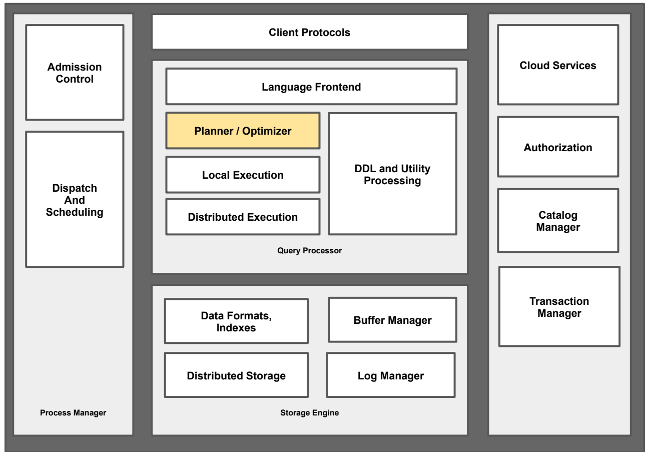
ZetaSQL

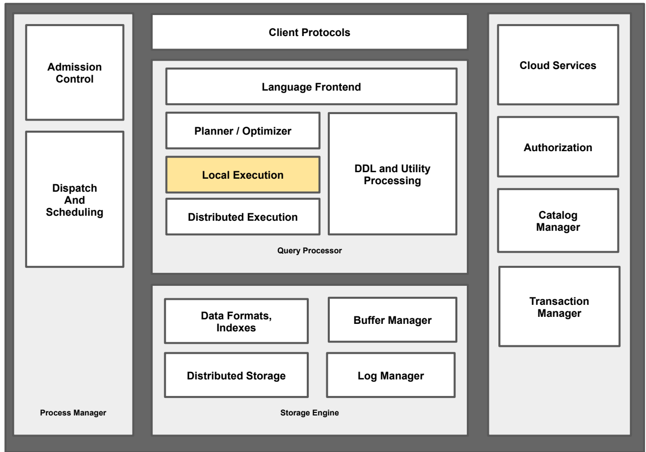


APACHE calcite™

Used by

Hive, beam, APACHE STORM™, NGDATA, APACHE DRILL, HerdDB, Apache Apex™, druid, Apache Kylin, APACHE PHOENIX, Apache Solr, Alibaba Group, Tencent, Apache Flink, samza, CASCADING, Qubole, ANT FINANCIAL, dremio, TBDS, omni sci, SQL Gremlin, UBER




Admission Control
Dispatch And Scheduling
Process Manager
Client Protocols
Language Frontend
Planner / Optimizer
Local Execution
Distributed Execution
DDL and Utility Processing
Query Processor
Data Formats, Indexes
Buffer Manager
Distributed Storage
Log Manager
Storage Engine
Cloud Services
Authorization
Catalog Manager
Transaction Manager

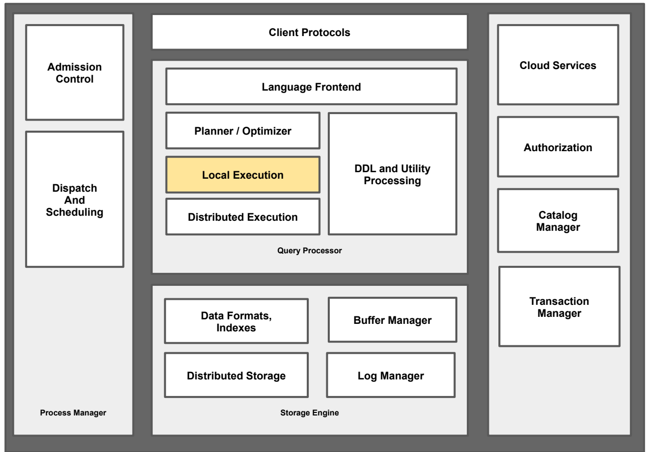
SQLite

DuckDB

DATA FUSION

Velox

ClickHouse



SQLite

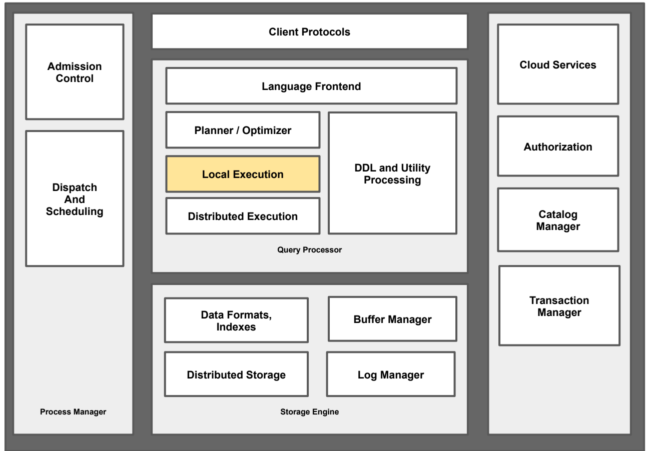
DuckDB

DATA FUSION

Velox

ClickHouse

ARROW



SQLite

DuckDB

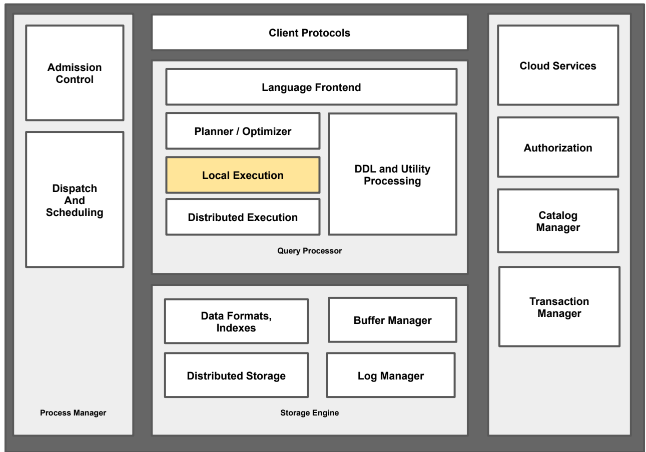
DATA FUSION

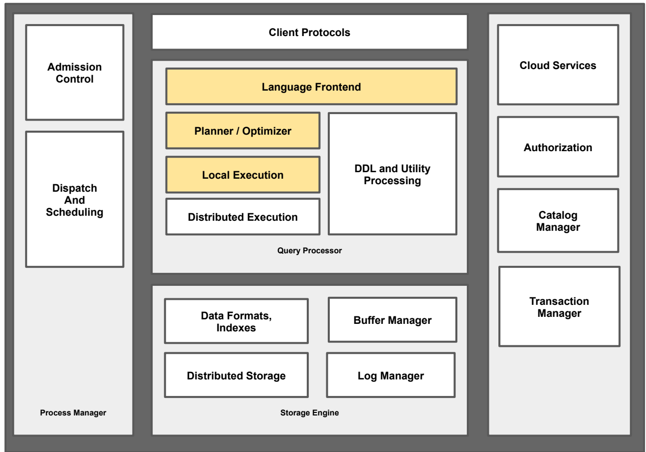
Velox

ClickHouse

ARROW

Unbundled Execution Engine ?




Admission Control
Dispatch And Scheduling
Process Manager
Client Protocols
Language Frontend
Planner / Optimizer
Local Execution
Distributed Execution
DDL and Utility Processing
Query Processor
Data Formats, Indexes
Buffer Manager
Distributed Storage
Log Manager
Storage Engine
Cloud Services
Authorization
Catalog Manager
Transaction Manager

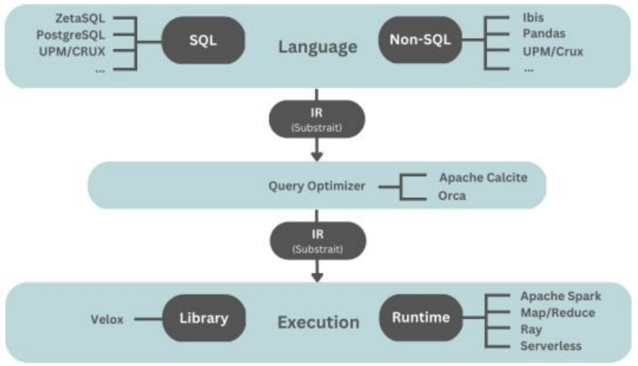
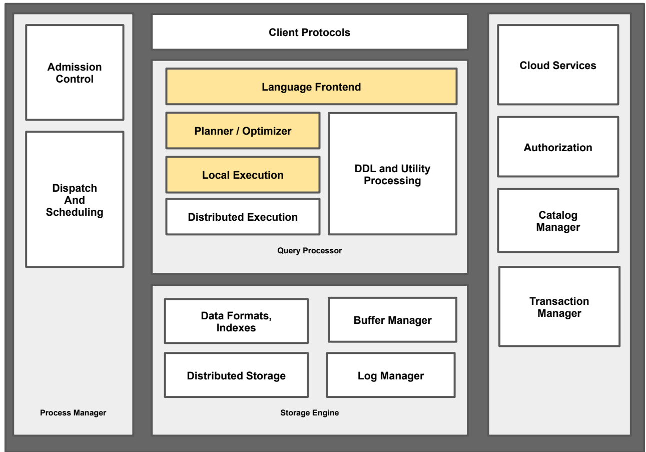


Figure 1: Open source modular data stack outline.

The Composable Data Management System Manifesto

https://dl.acm.org/doi/pdf/10.14778/3603581.3603604



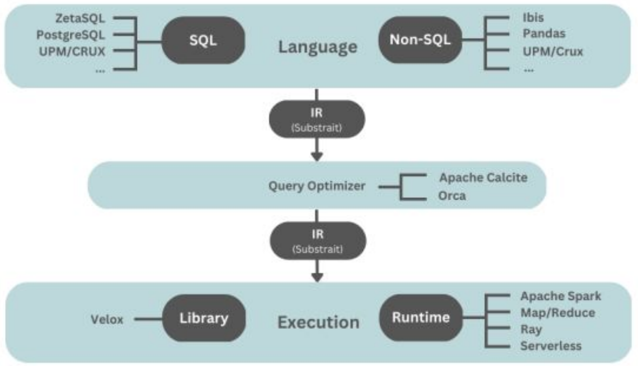
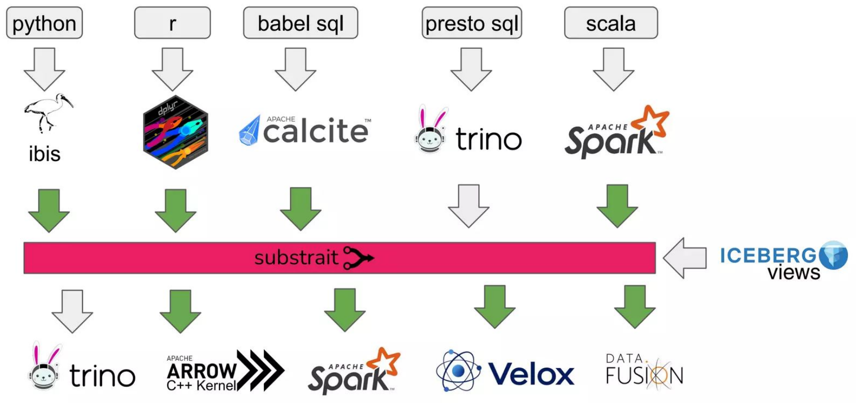
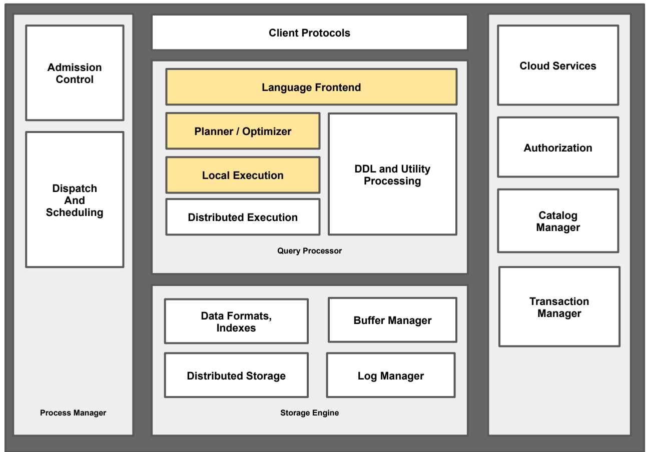


Figure 1: Open source modular data stack outline.

The Composable Data Management System Manifesto
https://dl.acm.org/doi/pdf/10.14778/3603581.3603604



Substrait: Rethinking DBMS Composability
https://cdmsworkshop.github.io/2022/Proceedings/Keynotes/Abstract_JacquesNadeau.pdf



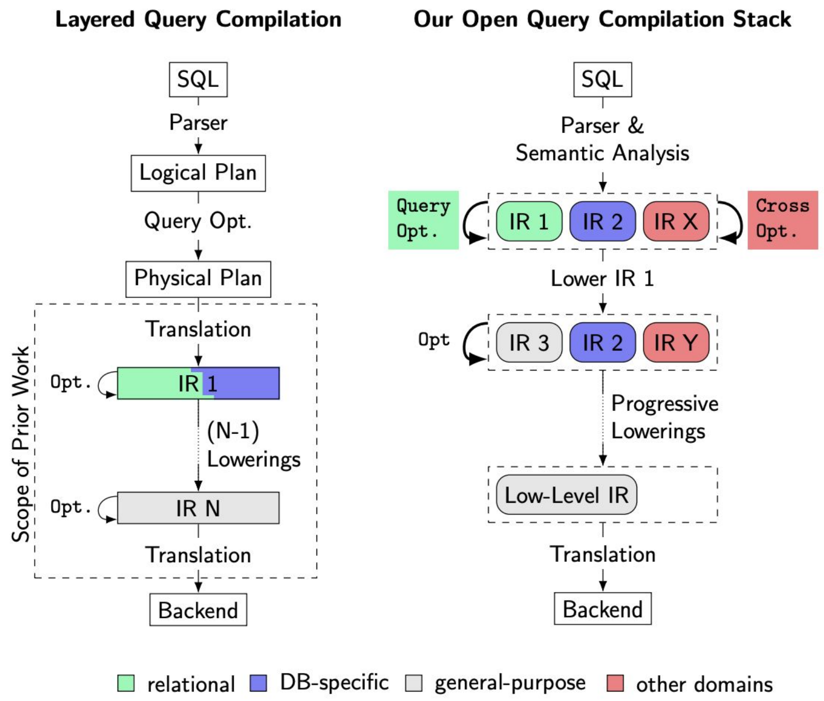
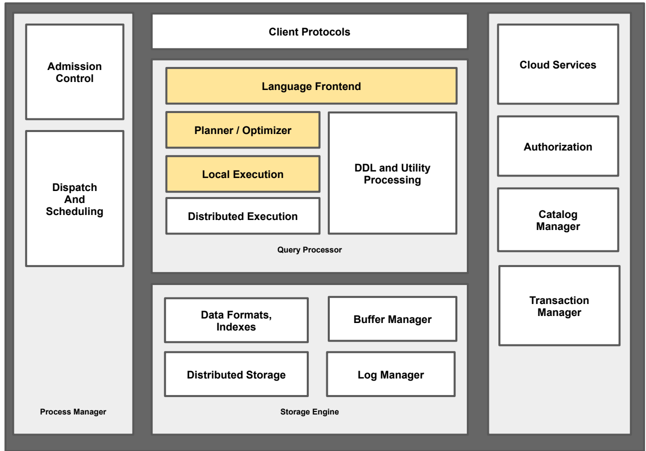
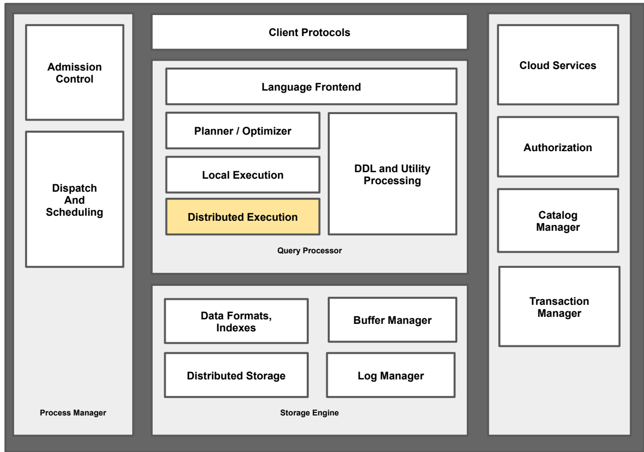
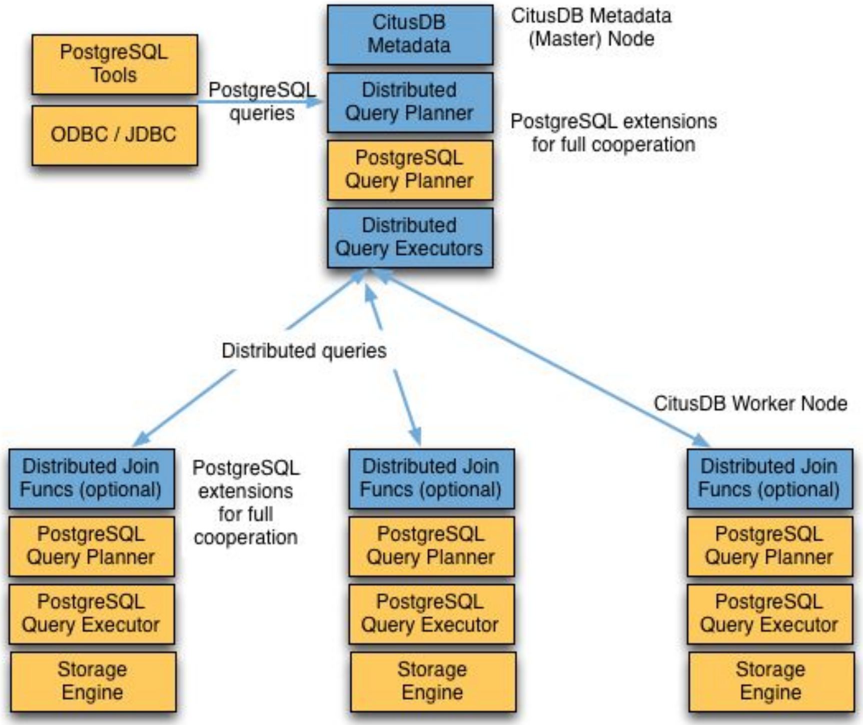


Figure 1: Our proposal of an *open* query compilation stack. It enhances prior work on layered query compilation with two major ideas: 1) Introducing open IRs, designed to be combined with other IRs, and 2) implementing query optimization as compiler passes.

Designing an Open Framework for Query Optimization and Compilation
https://db.in.tum.de/~jungmair/papers/p2485-jungmair.pdf

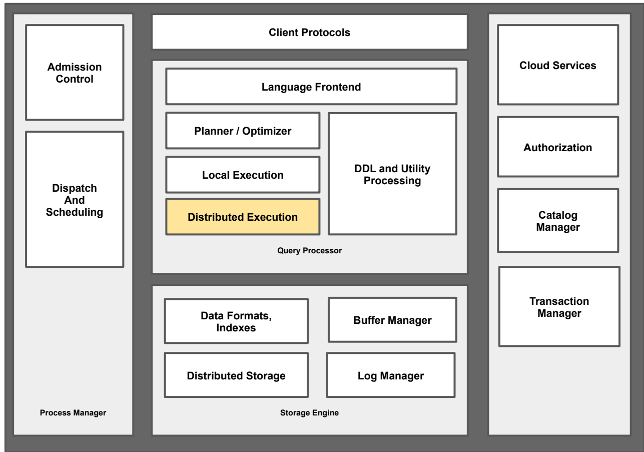



Admission Control
Dispatch And Scheduling
Process Manager
Client Protocols
Language Frontend
Planner / Optimizer
Local Execution
Distributed Execution
DDL and Utility Processing
Query Processor
Data Formats, Indexes
Buffer Manager
Distributed Storage
Log Manager
Storage Engine
Cloud Services
Authorization
Catalog Manager
Transaction Manager


PostgreSQL Tools
ODBC / JDBC
PostgreSQL queries
CitusDB Metadata
CitusDB Metadata (Master) Node
Distributed Query Planner
PostgreSQL extensions for full cooperation
PostgreSQL Query Planner
Distributed Query Executors
Distributed queries
CitusDB Worker Node
Distributed Join Funcs (optional)
PostgreSQL extensions for full cooperation
PostgreSQL Query Planner
PostgreSQL Query Executor
Storage Engine

Citus Query Processing

https://docs.citusdata.com/en/v7.0/performance/query_processing.html

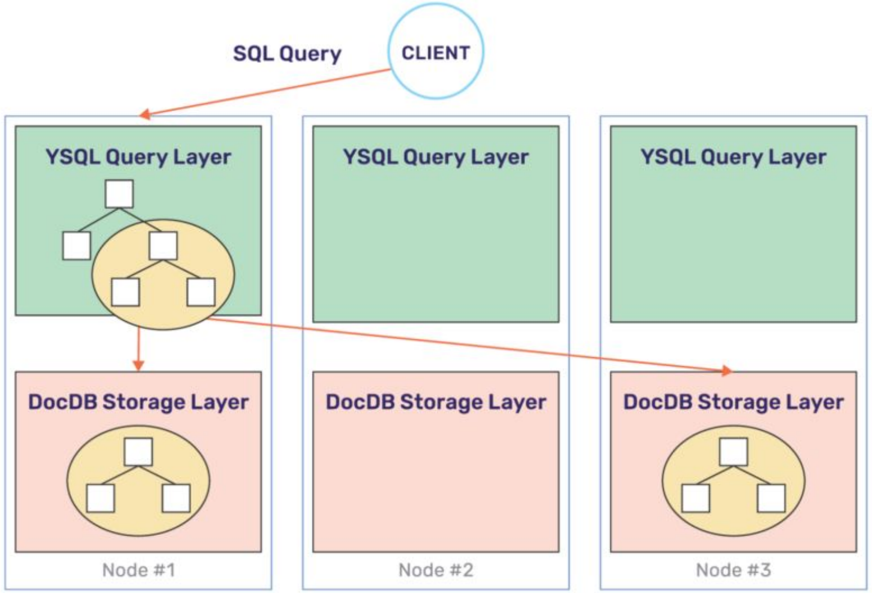


1. Build PostgreSQL execution tree

2. Identify subtree to pushdown

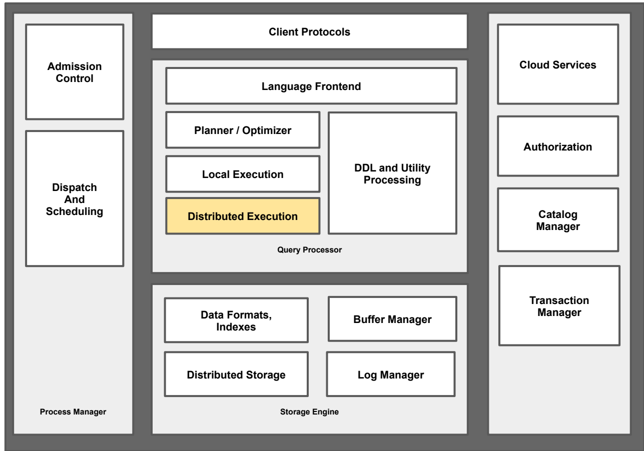
3. Identify nodes to send pushdown to

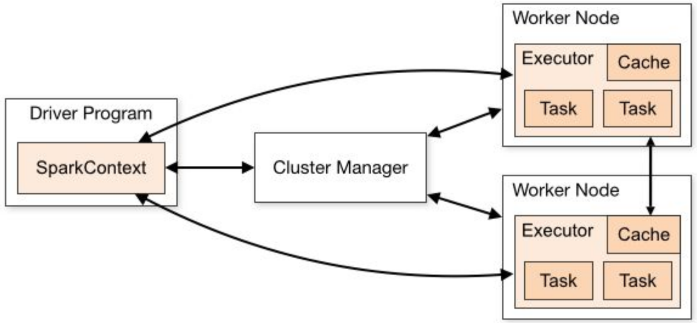
4. Execute subtree using PostgreSQL libraries on DocDB storage layer



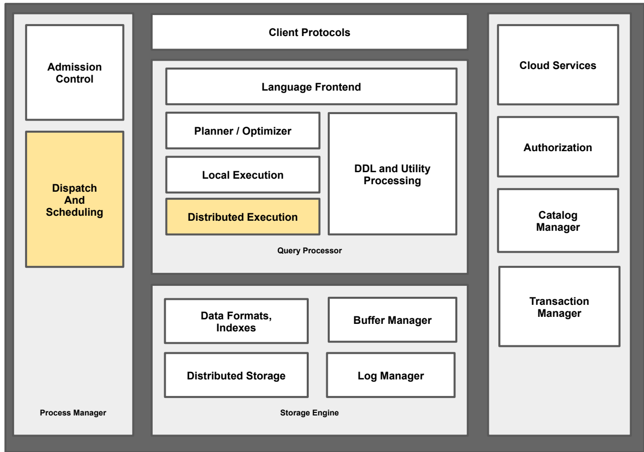
Generic pushdown mechanism in YugabyteDB

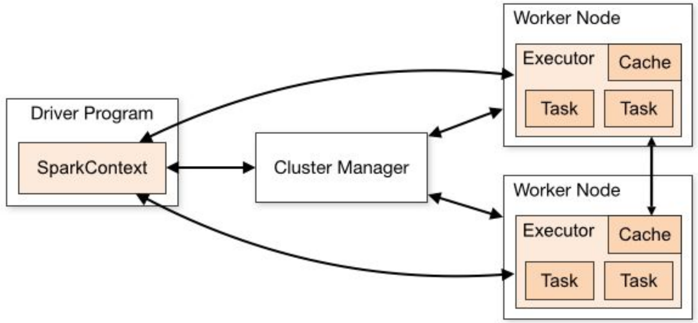
5 Query Pushdowns for Distributed SQL and How They Differ from a Traditional RDBMS
https://www.yugabyte.com/blog/5-query-pushdowns-for-distributed-sql-and-how-they-differ-from-a-traditional-rdbms/





Spark Cluster Mode Overview
https://spark.apache.org/docs/latest/cluster-overview.html





Spark Cluster Mode Overview
https://spark.apache.org/docs/latest/cluster-overview.html

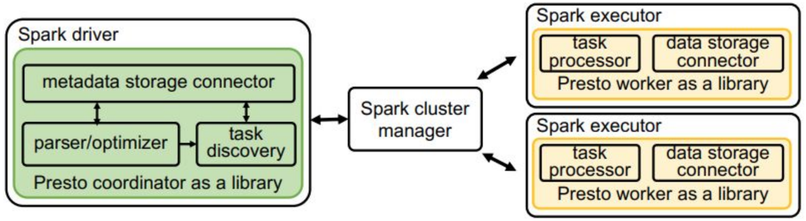
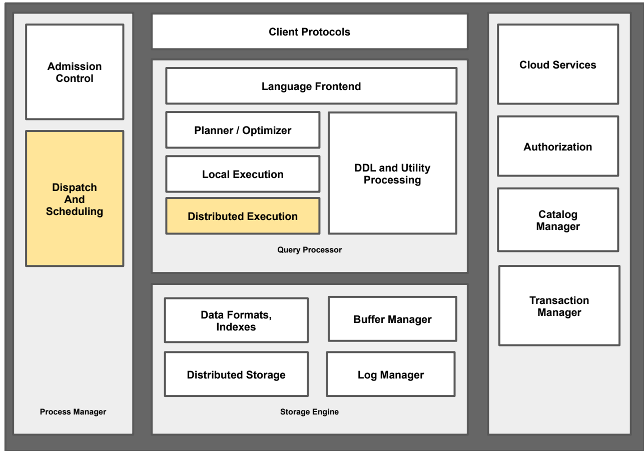
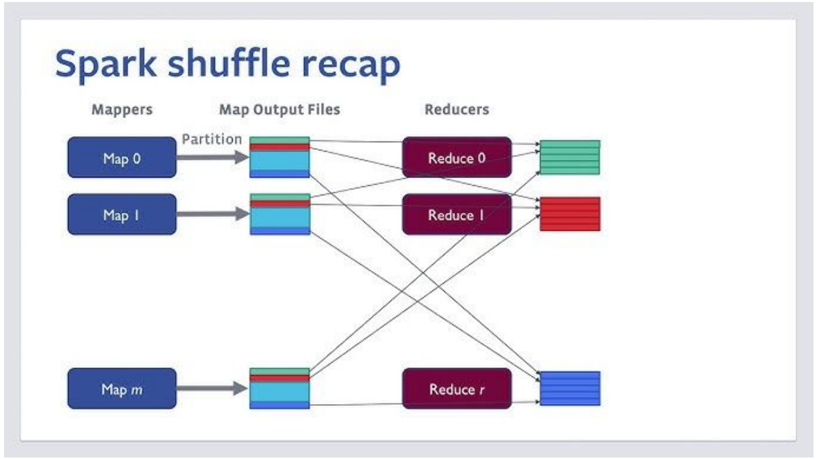
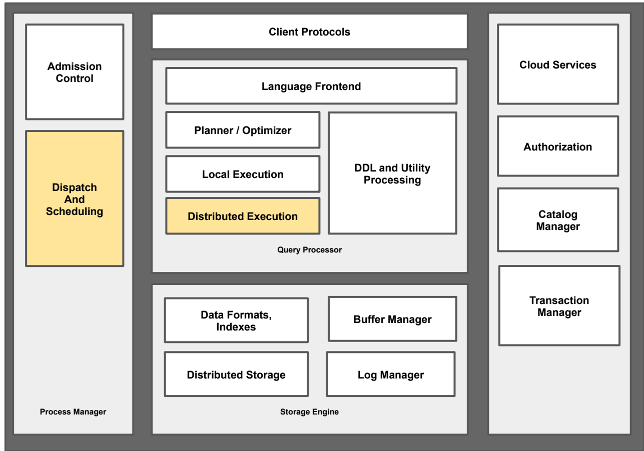


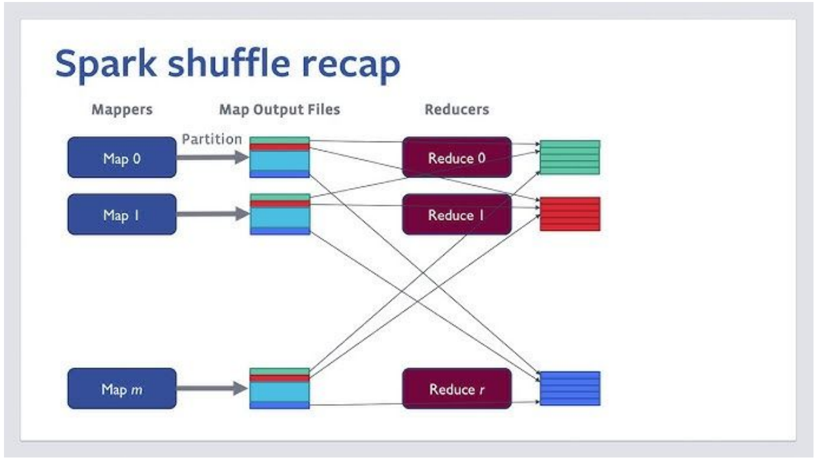
Figure 8: Presto on Spark architecture

Presto: A Decade of SQL Analytics at Meta
https://research.facebook.com/publications/presto-a-decade-of-sql-analytics-at-meta/

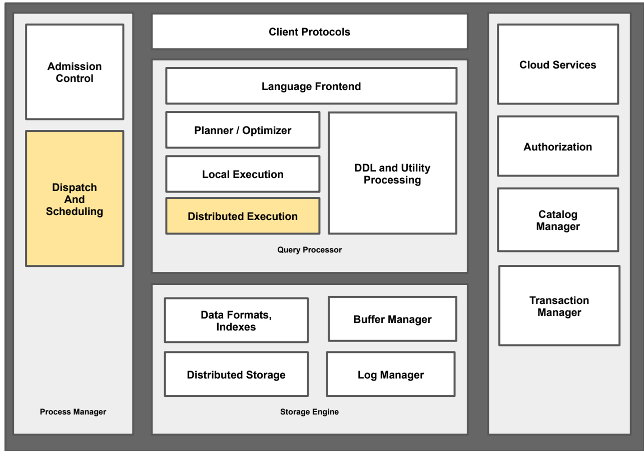



Spark shuffle recap





- InBuilt Shuffle service
- Standalone Shuffle Service
- YARN Shuffle Service
- Mesos Shuffle Service
- Kubernetes Shuffle Service
- Cosco (Meta)
- Magnet (Linkedin)
- Riffle (Meta)
- Zeus (Uber)
- EMR Remote Shuffle Service (Alibaba)



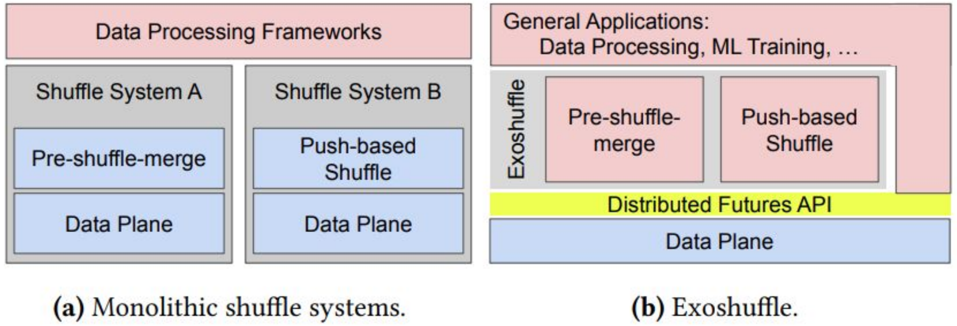
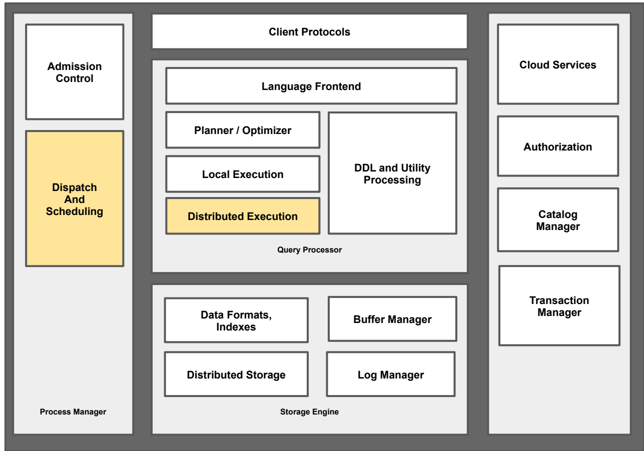
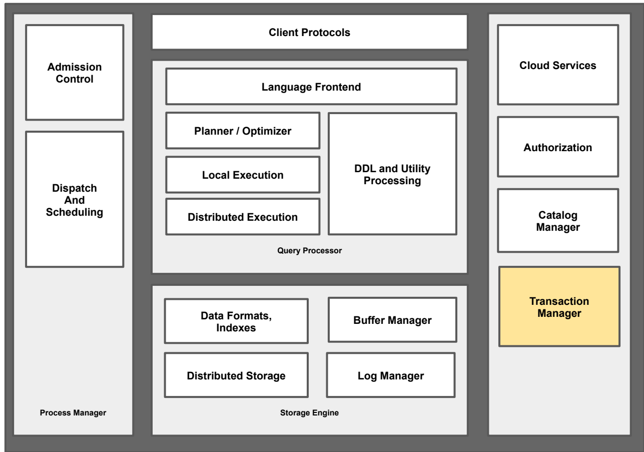


Figure 1: Exoshuffle builds on an extensible architecture. Shuffle as a library is easier to develop and more flexible to integrate with applications. The data plane ensures performance and reliability.

Exoshuffle: An Extensible Shuffle Architecture

https://arxiv.org/abs/2203.05072

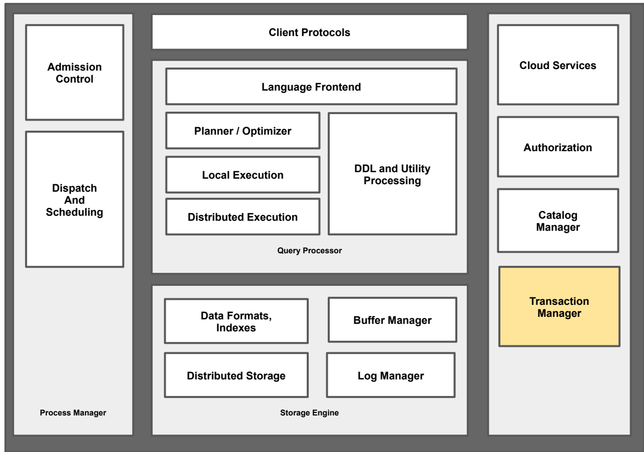



Admission Control
Dispatch And Scheduling
Process Manager
Client Protocols
Language Frontend
Planner / Optimizer
Local Execution
Distributed Execution
DDL and Utility Processing
Query Processor
Data Formats, Indexes
Buffer Manager
Distributed Storage
Log Manager
Storage Engine
Cloud Services
Authorization
Catalog Manager
Transaction Manager

DELTA LAKE

ICEBERG

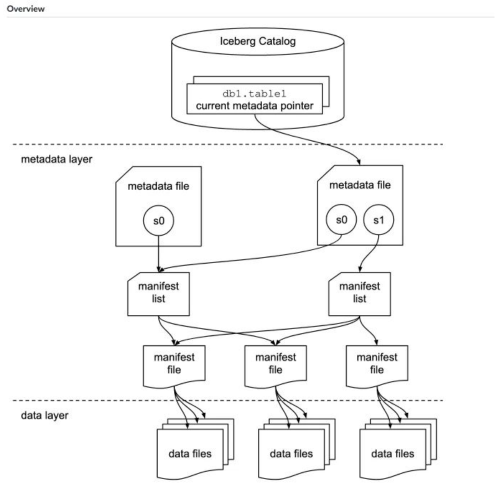
Apache hudi

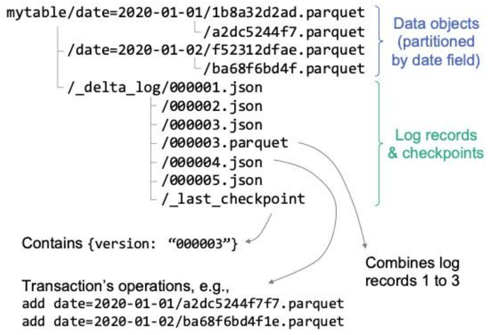


DELTA LAKE

ICEBERG

Apache hudi





```
mytable/date=2020-01-01/1b8a32d2ad.parquet
                       └/a2dc5244f7.parquet
       /date=2020-01-02/f52312dfae.parquet
                       └/ba68f6bd4f.parquet
       └/_delta_log/000001.json
                   /000002.json
                   /000003.json
                   /000003.parquet
                   /000004.json
                   /000005.json
                   └/_last_checkpoint
```

Data objects (partitioned by date field)

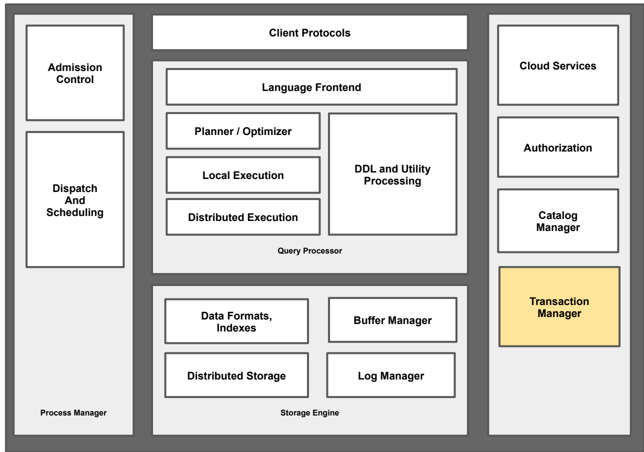
Log records & checkpoints

Contains {version: "000003"}

Transaction's operations, e.g.,

```
add date=2020-01-01/a2dc5244f7f7.parquet
add date=2020-01-02/ba68f6bd4f1e.parquet
```

Combines log records 1 to 3

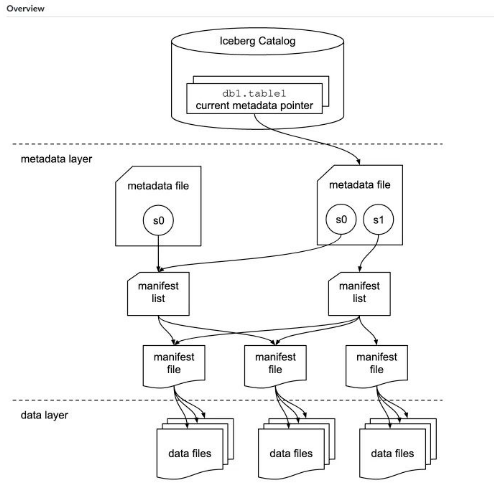


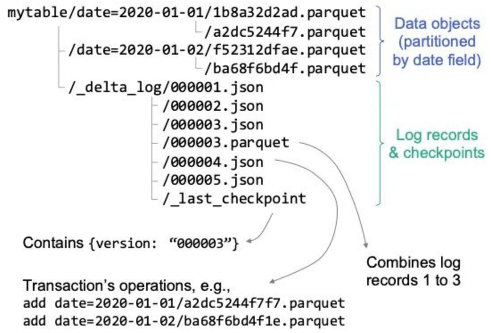
AWS Glue HIVE Nessie JDBC

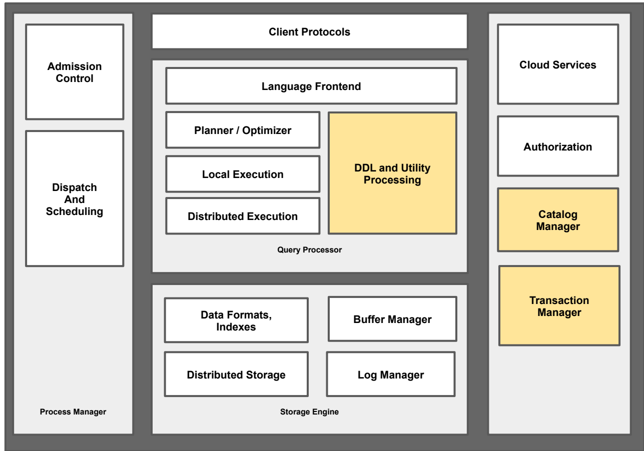
DELTA LAKE

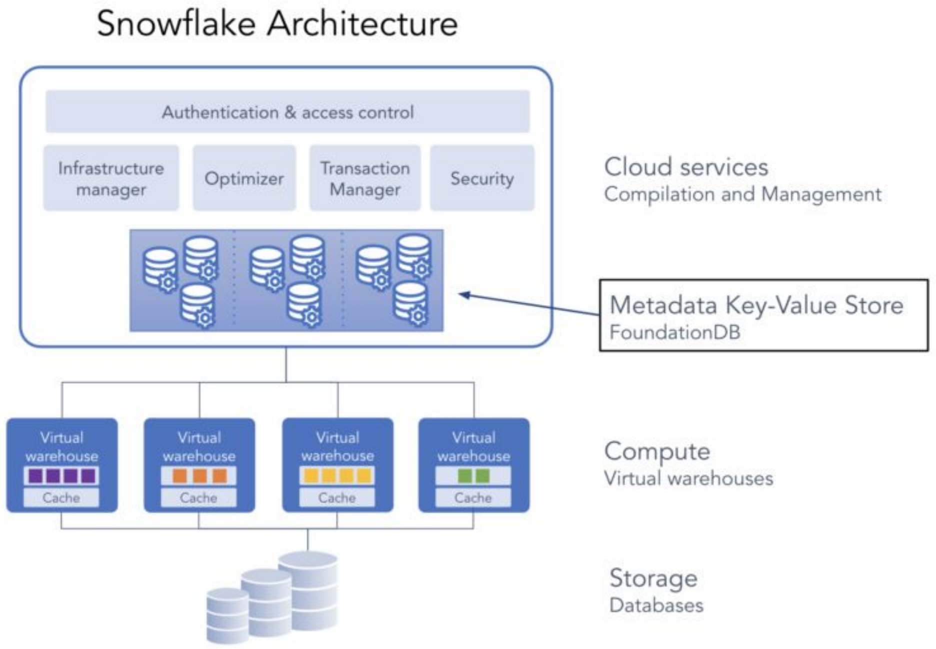
ICEBERG

Apache hudi

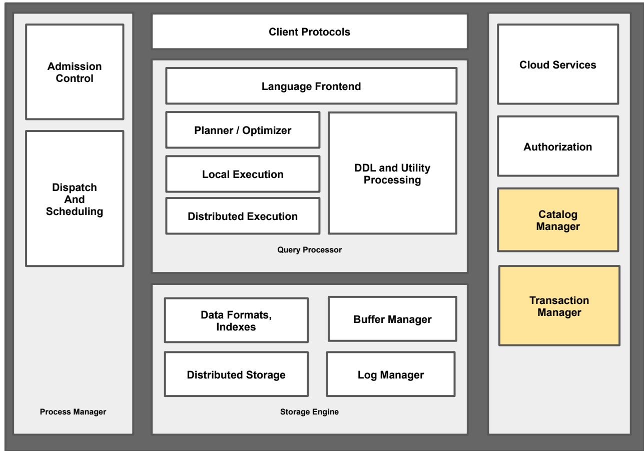


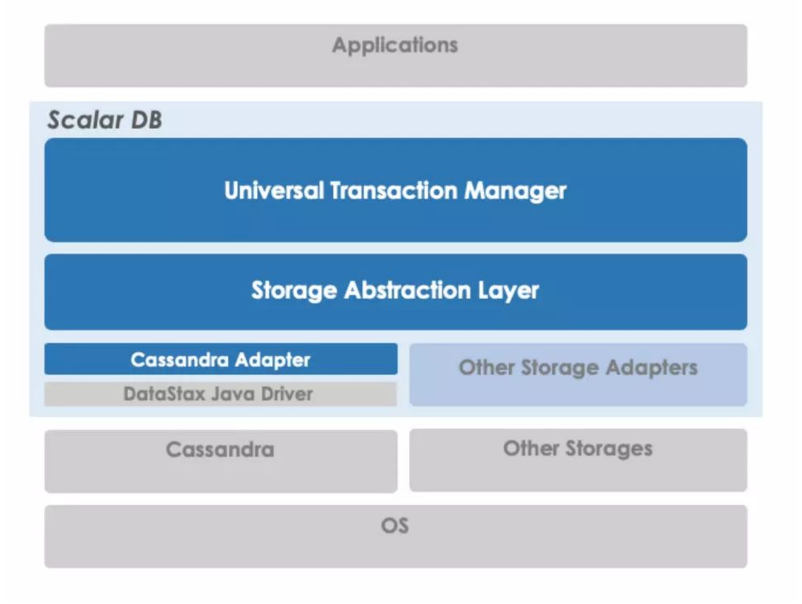

```
mytable/date=2020-01-01/1b8a32d2ad.parquet
                       └/a2dc5244f7.parquet
       /date=2020-01-02/f52312dfae.parquet
                       └/ba68f6bd4f.parquet
       └/_delta_log/000001.json
                   /000002.json
                   /000003.json
                   /000003.parquet
                   /000004.json
                   /000005.json
                   └/_last_checkpoint
```
Data objects (partitioned by date field)
Log records & checkpoints
Contains {version: "000003"}
Transaction's operations, e.g.,
add date=2020-01-01/a2dc5244f7f7.parquet
add date=2020-01-02/ba68f6bd4f1e.parquet
Combines log records 1 to 3



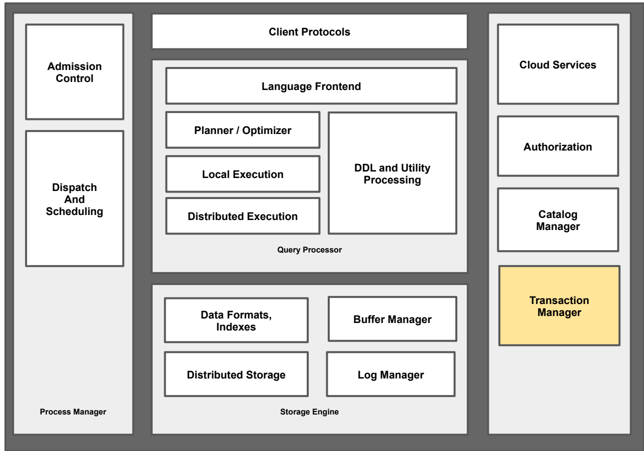


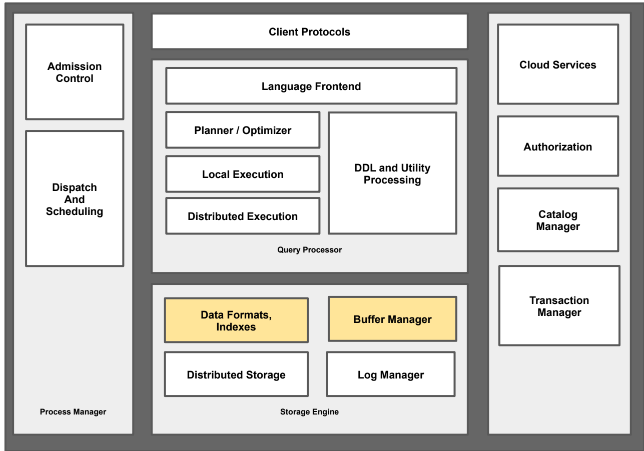
FOUNDATIONDB



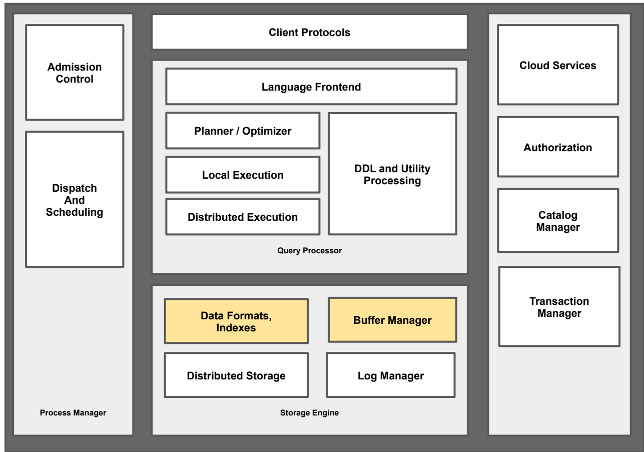


ScalarDB: Universal Transaction Manager for Polystores
https://github.com/scalar-labs/scalardb




Admission Control
Dispatch And Scheduling
Process Manager
Client Protocols
Language Frontend
Planner / Optimizer
Local Execution
Distributed Execution
DDL and Utility Processing
Query Processor
Data Formats, Indexes
Buffer Manager
Distributed Storage
Log Manager
Storage Engine
Cloud Services
Authorization
Catalog Manager
Transaction Manager

RocksDB	26
levelDB	10
Pebble	1
SQLite	13
LMDB	12



RocksDB	26
levelDB	10
Pebble	1
SQLite	13
LMDB	12

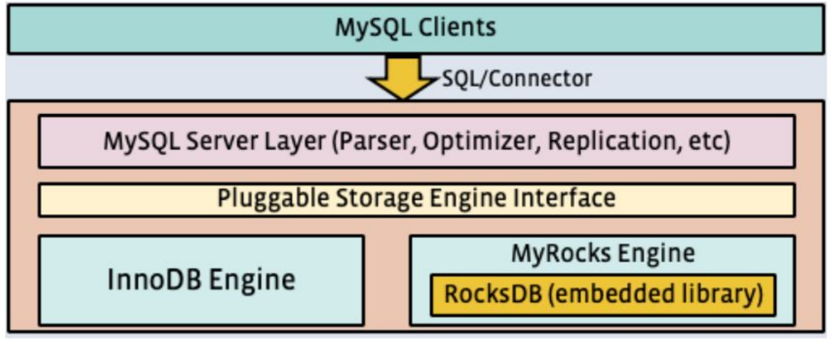
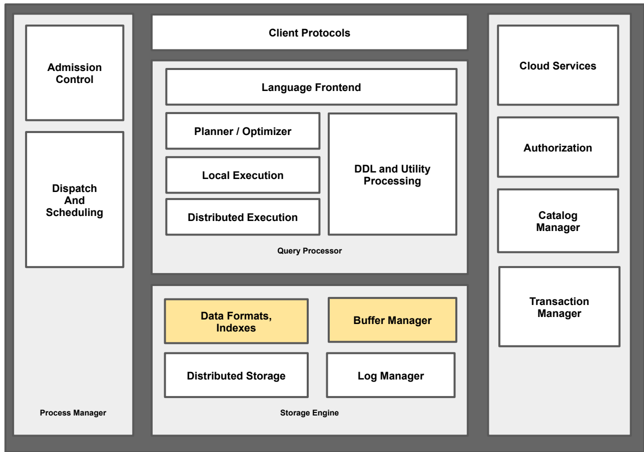

MySQL Clients
SQL/Connector
MySQL Server Layer (Parser, Optimizer, Replication, etc)
Pluggable Storage Engine Interface
InnoDB Engine
MyRocks Engine
RocksDB (embedded library)

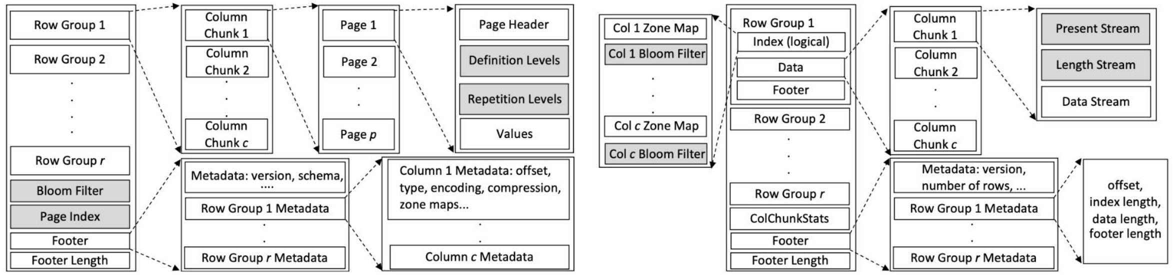
Figure 1: MySQL and MyRocks Storage Engine

MyRocks: LSM-Tree Database Storage Engine Serving Facebook's Social Graph
https://www.vldb.org/pvldb/vol13/p3217-matsunobu.pdf



		Parquet	ORC
Features	Internal Layout (§3.1)	PAX	PAX
	Encoding Variants (§3.2)	plain, RLE_DICT, RLE, Delta, Bitpacking	plain, RLE_DICT, RLE, Delta, Bitpacking, FOR
	Compression (§3.3)	Snappy, gzip, LZO, zstd, LZ4, Brotli	Snappy, zlib, LZO, zstd, LZ4
	Type System (§3.4)	Separate logical and physical type system	One unified type system
	Zone Map / Index (§3.5)	Min-max per smallest zone map/row group/file	Min-max per smallest zone map/row group/file
	Bloom Filter (§3.5)	Supported per column chunk	Supported per smallest zone map
	Nested Data Encoding (§3.6)	Dremel Model	Length and presence
Concepts	Row Group	Row Group	Stripe
	Smallest Zone Map	Page Index (a Page)	Row Index (10k rows)
	Encoding Unit	Page	Stream
	Compression Unit	Page	Compression Chunk

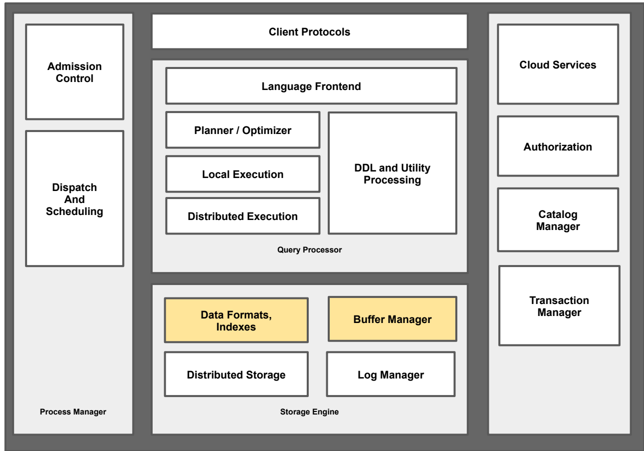
Table 1: Feature Taxonomy and Concepts Mapping

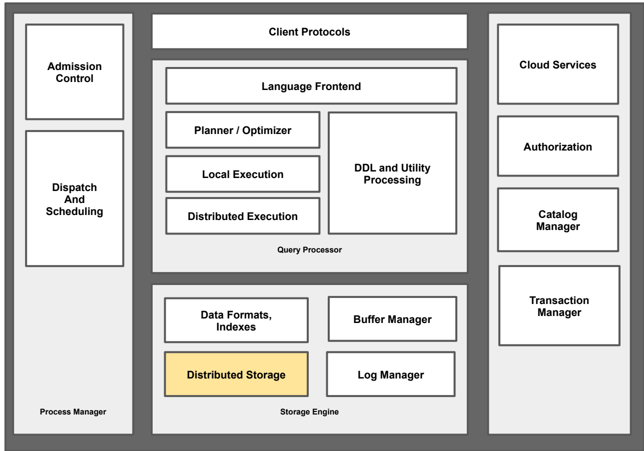


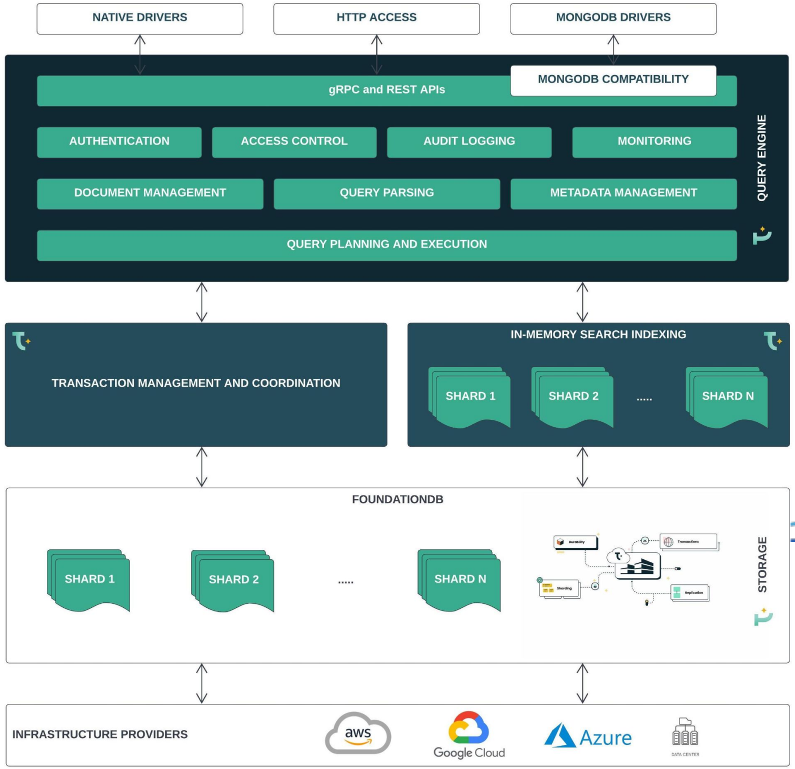
(a) Parquet layout. **(b)** ORC layout.

Figure 1: Illustration of file layout — Parquet (a) and ORC (b). Blocks in gray are optional depending on configurations/data.

An Empirical Evaluation of Columnar Storage Formats
https://arxiv.org/pdf/2304.05028.pdf

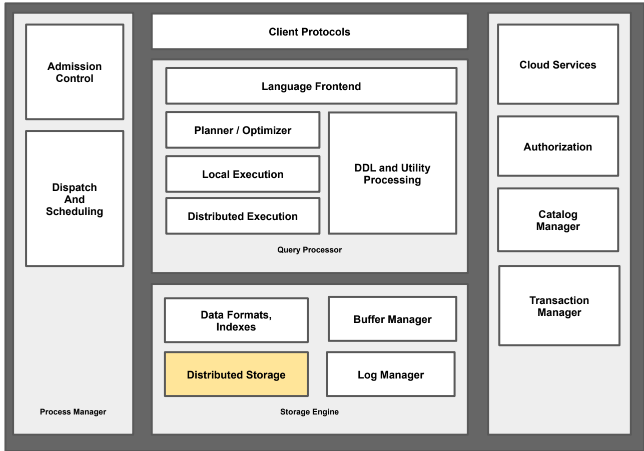



Admission Control
Dispatch And Scheduling
Process Manager
Client Protocols
Language Frontend
Planner / Optimizer
Local Execution
Distributed Execution
DDL and Utility Processing
Query Processor
Data Formats, Indexes
Buffer Manager
Distributed Storage
Log Manager
Storage Engine
Cloud Services
Authorization
Catalog Manager
Transaction Manager



Tigris - Document Database built on FoundationDB

https://www.tigrisdata.com/docs/concepts/architecture/



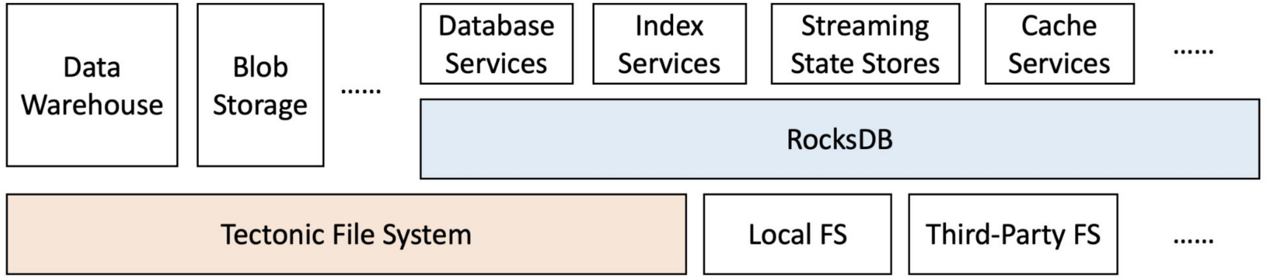
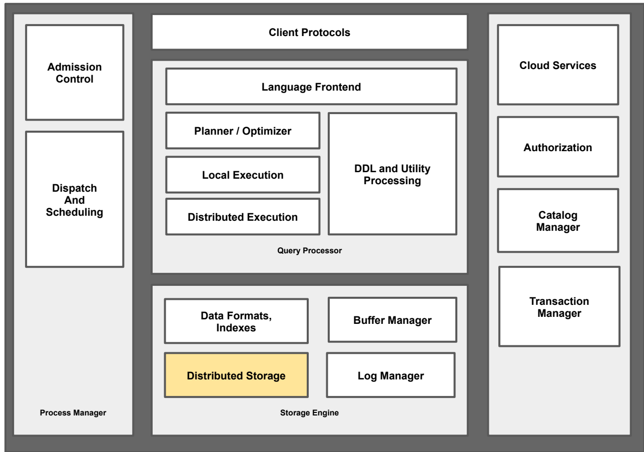
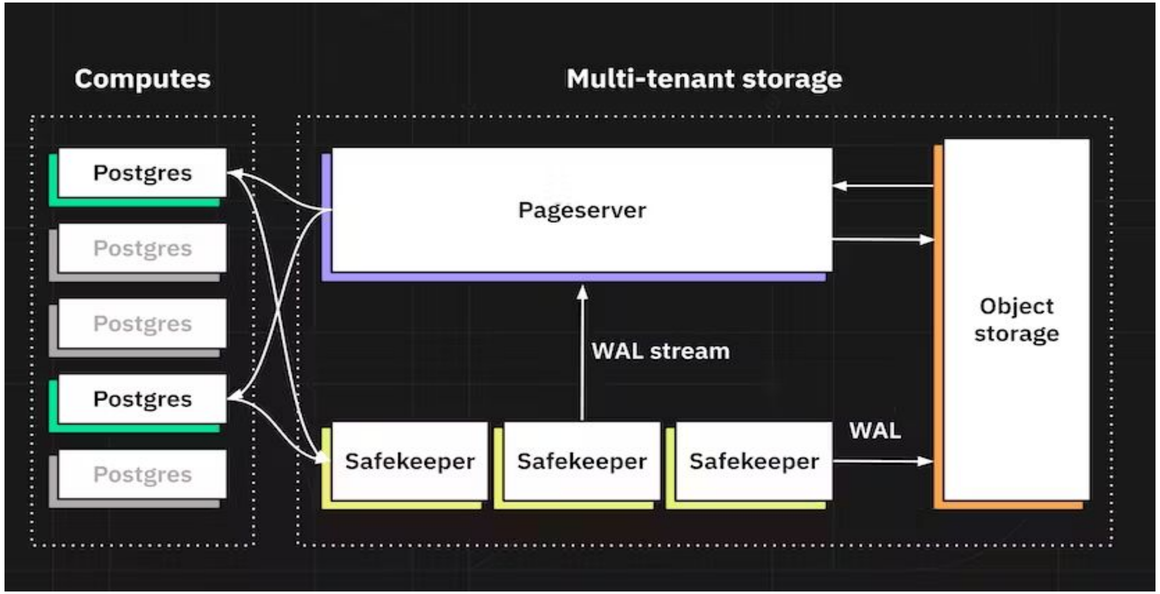


Fig. 1. Different applications are able to run on disaggregated storage with RocksDB.

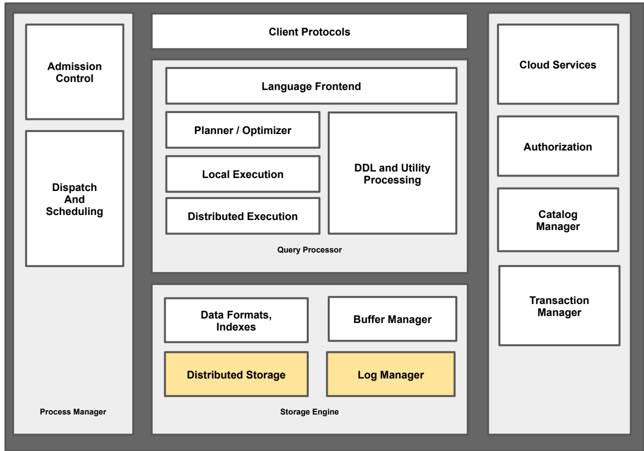
Disaggregating RocksDB: A Production Experience

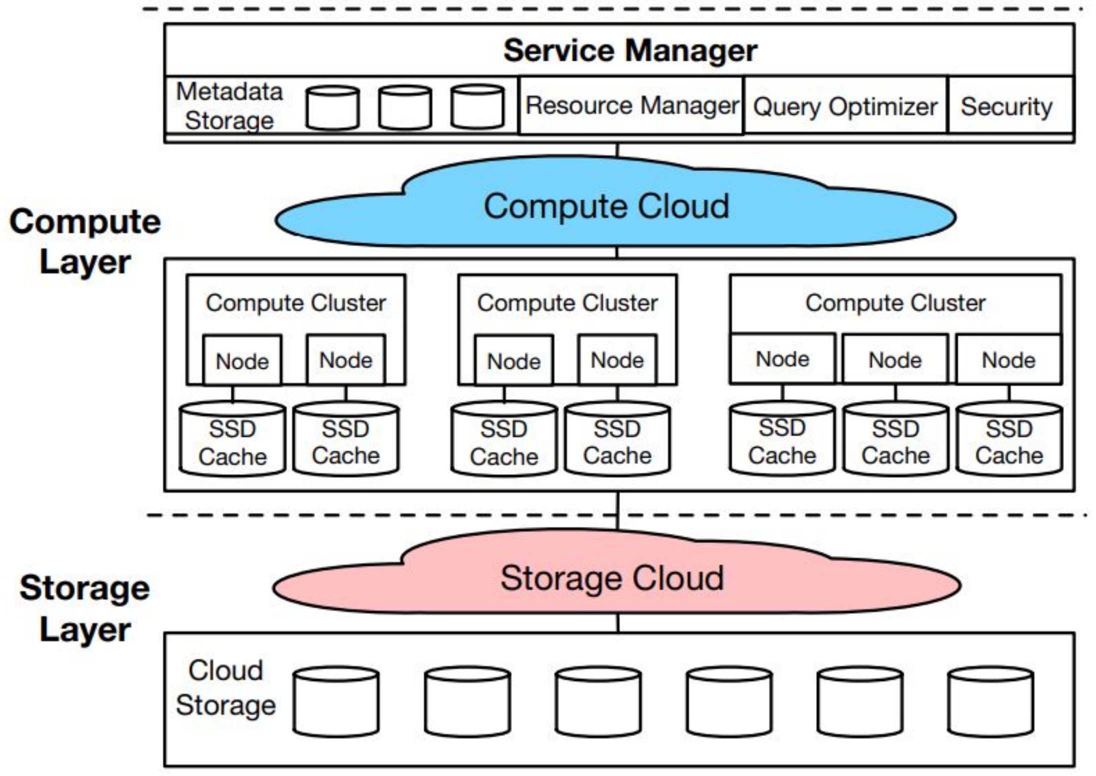
https://research.facebook.com/publications/disaggregating-rocksdb-a-production-experience/



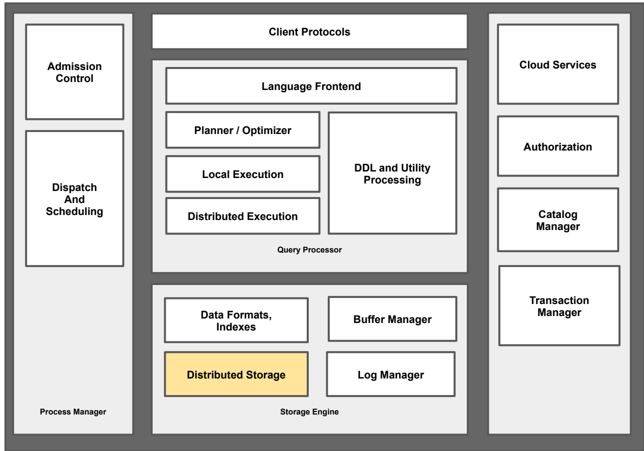


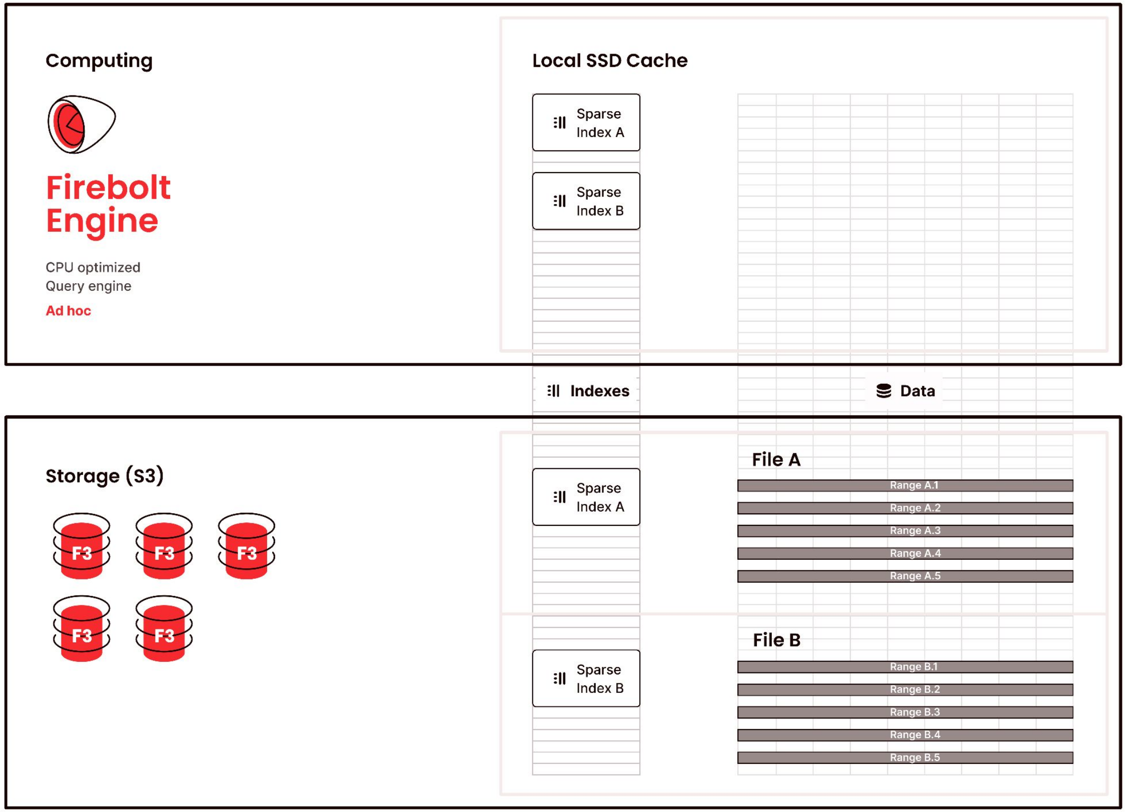
Neon Architecture
https://neon.tech/docs/introduction/architecture-overview

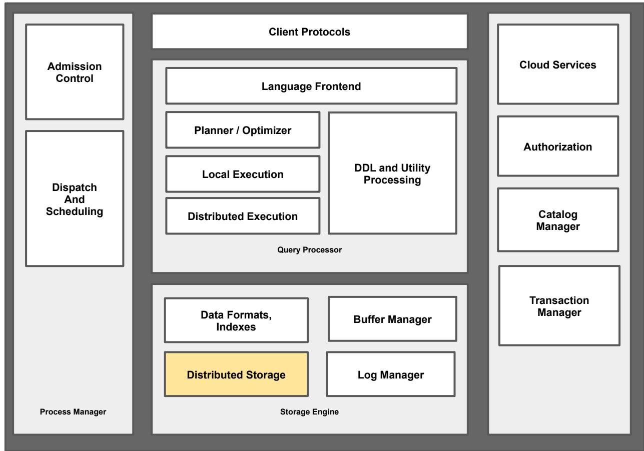


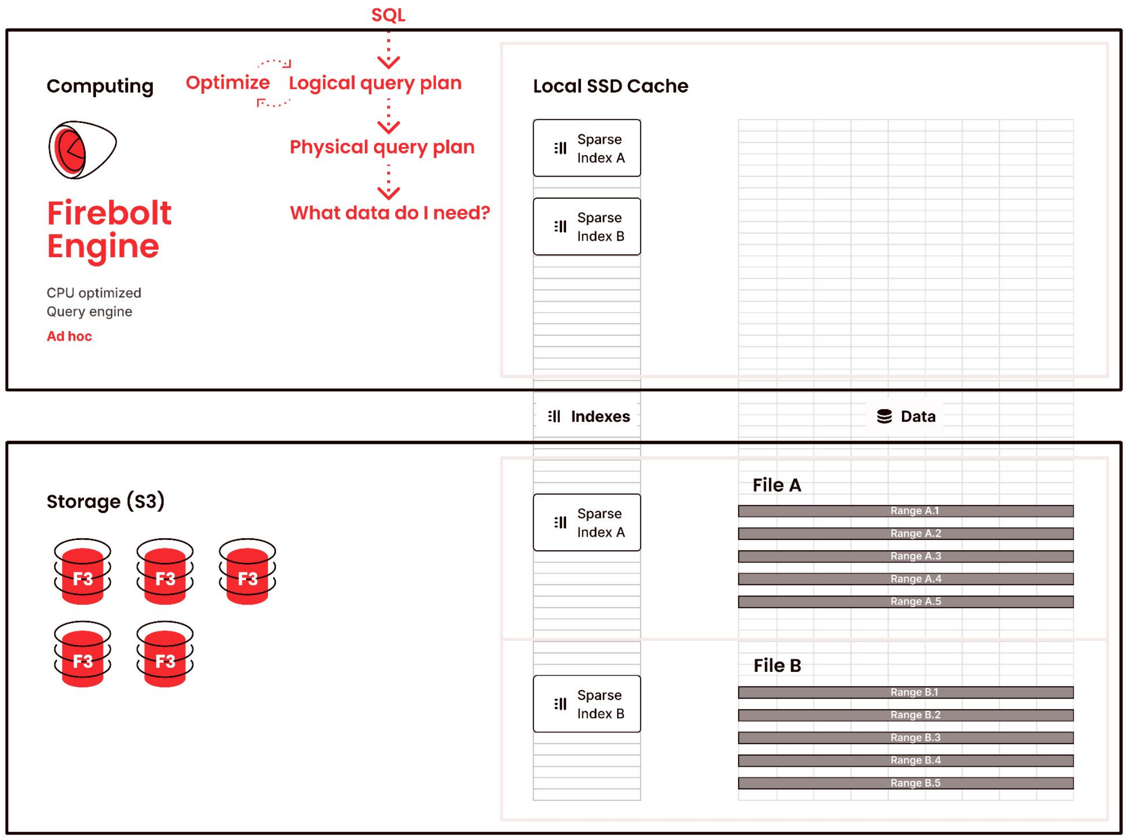


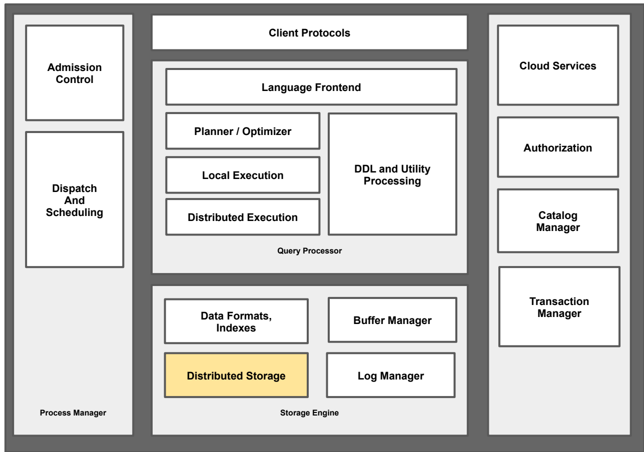
(1) Disaggregated Compute-Storage OLAP Architecture

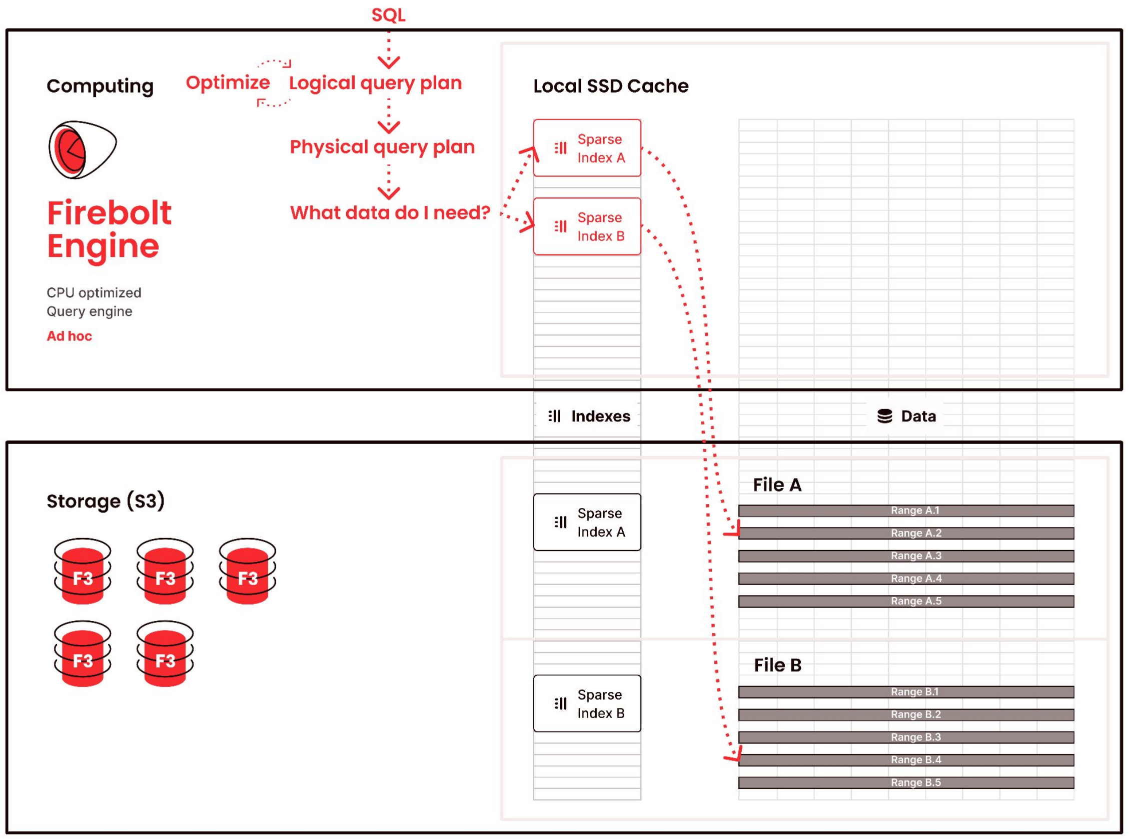


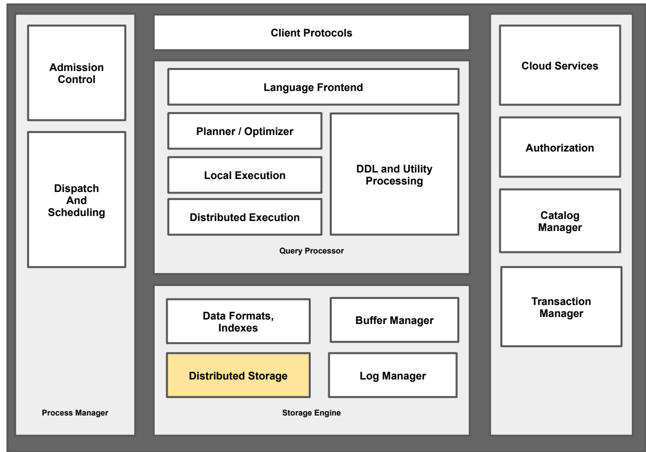

Computing
Firebolt Engine
CPU optimized
Query engine
Ad hoc
Local SSD Cache
Sparse Index A
Sparse Index B
Indexes
Data
Storage (S3)
F3
F3
F3
F3
F3
File A
Range A.1
Range A.2
Range A.3
Range A.4
Range A.5
File B
Range B.1
Range B.2
Range B.3
Range B.4
Range B.5
Sparse Index A
Sparse Index B

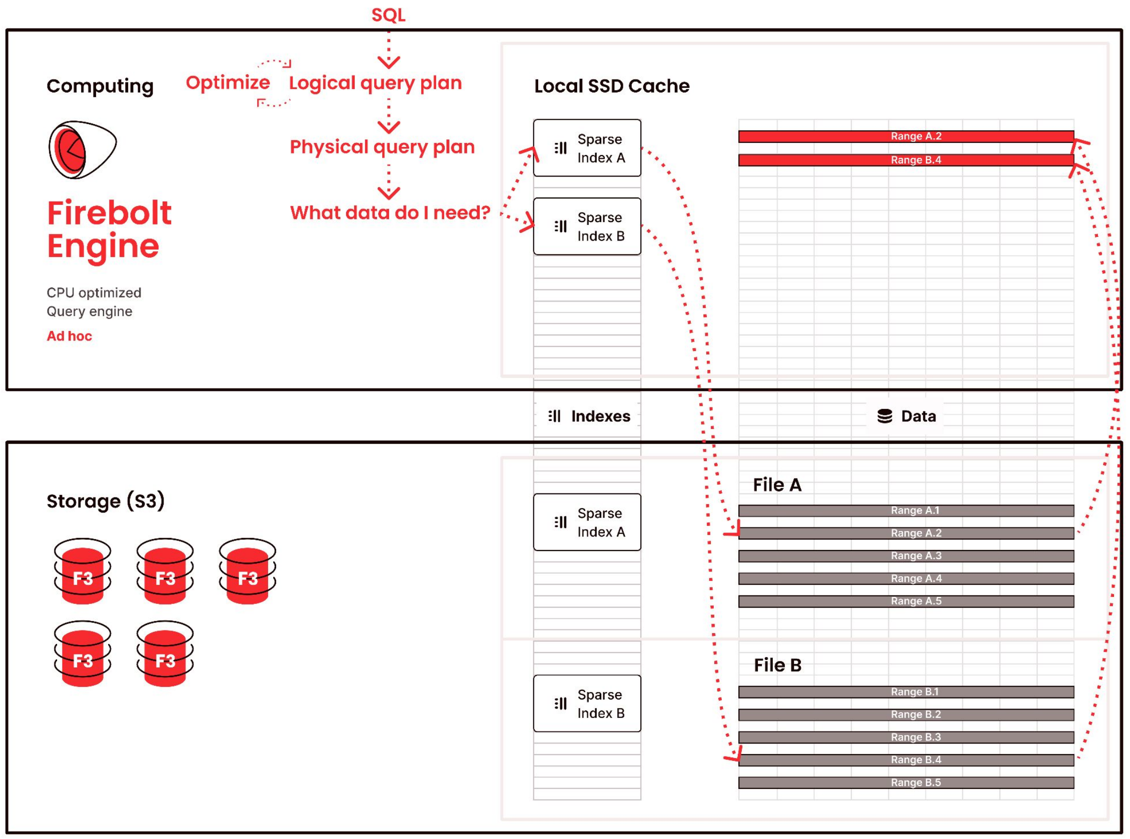

Admission Control
Dispatch And Scheduling
Process Manager
Client Protocols
Language Frontend
Planner / Optimizer
Local Execution
Distributed Execution
DDL and Utility Processing
Query Processor
Data Formats, Indexes
Buffer Manager
Distributed Storage
Log Manager
Storage Engine
Cloud Services
Authorization
Catalog Manager
Transaction Manager

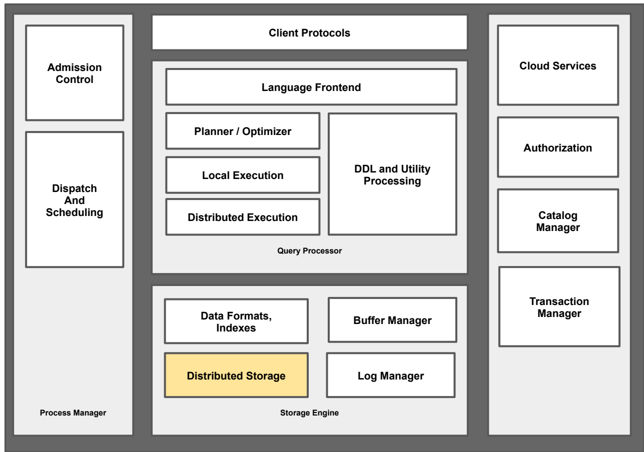

SQL
Computing
Optimize
Logical query plan
Physical query plan
What data do I need?
Firebolt Engine
CPU optimized
Query engine
Ad hoc
Local SSD Cache
Sparse Index A
Sparse Index B
Indexes
Data
Storage (S3)
F3
F3
F3
F3
F3
Sparse Index A
Sparse Index B
File A
Range A.1
Range A.2
Range A.3
Range A.4
Range A.5
File B
Range B.1
Range B.2
Range B.3
Range B.4
Range B.5

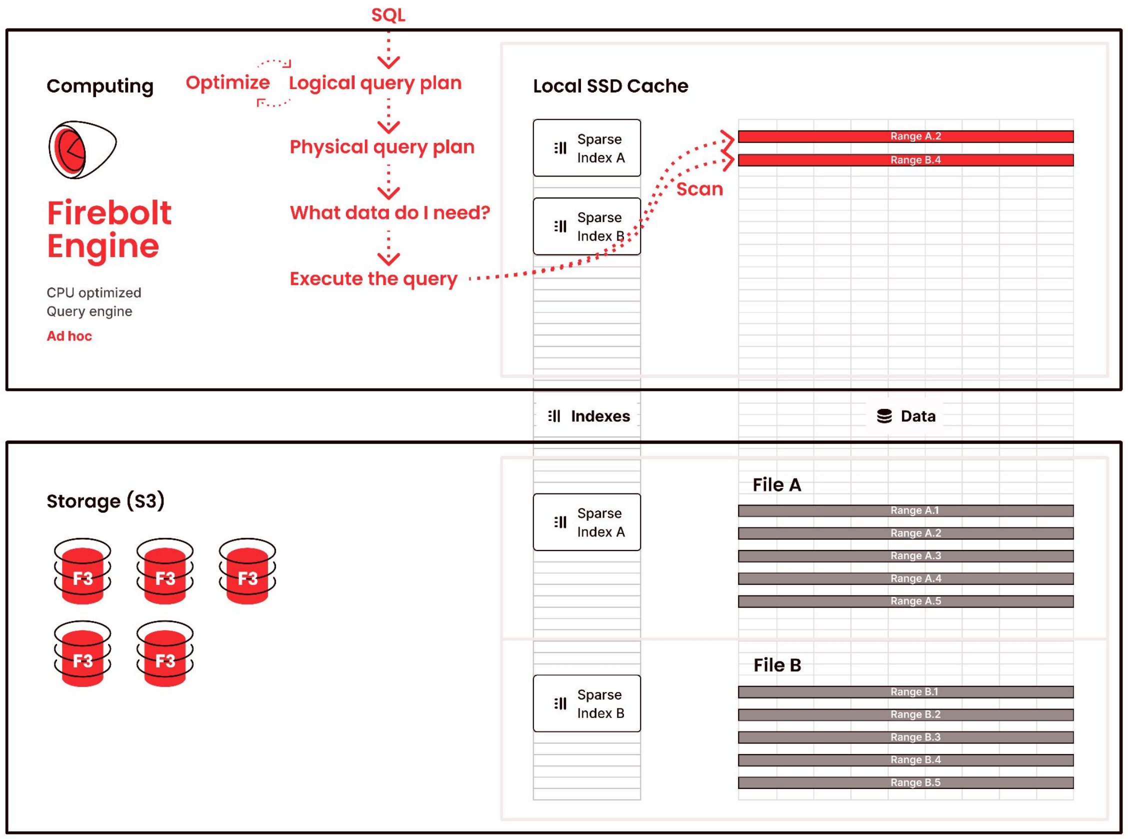

Admission Control
Dispatch And Scheduling
Process Manager
Client Protocols
Language Frontend
Planner / Optimizer
Local Execution
Distributed Execution
DDL and Utility Processing
Query Processor
Data Formats, Indexes
Buffer Manager
Distributed Storage
Log Manager
Storage Engine
Cloud Services
Authorization
Catalog Manager
Transaction Manager

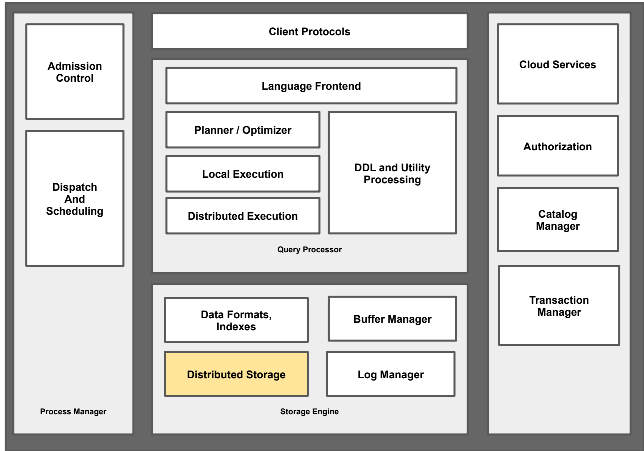

SQL
Computing
Optimize
Logical query plan
Physical query plan
What data do I need?
Firebolt Engine
CPU optimized
Query engine
Ad hoc
Local SSD Cache
Sparse Index A
Sparse Index B
Indexes
Data
Storage (S3)
F3
F3
F3
F3
F3
Sparse Index A
Sparse Index B
File A
Range A.1
Range A.2
Range A.3
Range A.4
Range A.5
File B
Range B.1
Range B.2
Range B.3
Range B.4
Range B.5


Admission Control
Dispatch And Scheduling
Process Manager
Client Protocols
Language Frontend
Planner / Optimizer
Local Execution
Distributed Execution
DDL and Utility Processing
Query Processor
Data Formats, Indexes
Buffer Manager
Distributed Storage
Log Manager
Storage Engine
Cloud Services
Authorization
Catalog Manager
Transaction Manager


SQL
Computing
Optimize
Logical query plan
Physical query plan
What data do I need?
Firebolt Engine
CPU optimized
Query engine
Ad hoc
Local SSD Cache
Sparse Index A
Sparse Index B
Range A.2
Range B.4
Indexes
Data
Storage (S3)
F3
File A
Range A.1
Range A.2
Range A.3
Range A.4
Range A.5
File B
Range B.1
Range B.2
Range B.3
Range B.4
Range B.5


Admission Control
Dispatch And Scheduling
Process Manager
Client Protocols
Language Frontend
Planner / Optimizer
Local Execution
Distributed Execution
DDL and Utility Processing
Query Processor
Data Formats, Indexes
Buffer Manager
Distributed Storage
Log Manager
Storage Engine
Cloud Services
Authorization
Catalog Manager
Transaction Manager


SQL
Computing
Optimize
Logical query plan
Physical query plan
What data do I need?
Execute the query
Firebolt Engine
CPU optimized
Query engine
Ad hoc
Local SSD Cache
Sparse Index A
Sparse Index B
Scan
Range A.2
Range B.4
Indexes
Data
Storage (S3)
F3
File A
Range A.1
Range A.2
Range A.3
Range A.4
Range A.5
File B
Range B.1
Range B.2
Range B.3
Range B.4
Range B.5


Client Protocols
Admission Control
Dispatch And Scheduling
Process Manager
Language Frontend
Planner / Optimizer
Local Execution
Distributed Execution
DDL and Utility Processing
Query Processor
Data Formats, Indexes
Buffer Manager
Distributed Storage
Log Manager
Storage Engine
Cloud Services
Authorization
Catalog Manager
Transaction Manager

	SELECT	DDL/DML/DCL/TCL	Total	
SQLite	7.2M	225K	7.4M	12K

	SELECT	DDL/DML/DCL/TCL	Total	
SQLite	7.2M	225K	7.4M	12K
	24K	22K	46K	4K

	SELECT	DDL/DML/DCL/TCL	Total	
SQLite	7.2M	225K	7.4M	12K
	24K	22K	46K	4K
	17K	28K	35K	3K

	SELECT	DDL/DML/DCL/TCL	Total	F
SQLite	7.2M	225K	7.4M	12K
	24K	22K	46K	4K
	17K	28K	35K	3K
G	45K	4K	49K	3K

	SELECT	DDL/DML/DCL/TCL	Total	F
SQLite	7.2M	225K	7.4M	12K
	24K	22K	46K	4K
	17K	28K	35K	3K
G	45K	4K	49K	3K
F	30K	7K	37K	22K