2018

upsolver

In-Memory Data Preparation

White Paper

Upsolver Overview

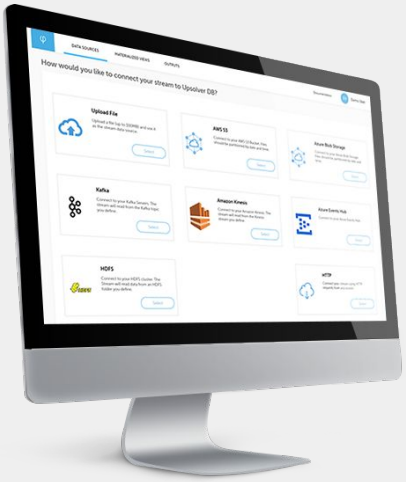
Index

The Big Data Preparation Nightmare

Companies strive to improve their products and services with big data coming from click streams, user activity, IoT, security logs and more.

To prepare such big data for analysis, companies spend months writing code, tuning performance and integrating several data platforms together. **Big data projects are notorious for being long, risky and expensive**.

Upsolver solves this problem with an easy-to-use cloud product for a scale of GBs to PBs. Data analysts, data scientists, product managers and developers use Upsolver to create analytics over their streaming data. They do it in hours, at low cost and without ongoing maintenance.



In-Memory

Data Preparation Platform

What's Upsolver

Upsolver is an In-Memory Data Preparation Platform.
It enables data professionals to prepare and deliver data, at a massive scale (10s of TBs to PBs), in a matter of minutes.

What makes Upsolver Unique

Upsolver shortens the implementation time of Big and Streaming Data projects from weeks/months to minutes, literally.

Fast - Big Data processing and delivery 10x faster than Spark

Simple - Configured in minutes with a drag & drop UI or SQL

Scalable - From 10s of GBs to PBs with Amazon S3 and EC2

The Underlying Technology

Powered by cutting edge Volcano™ technology, Upsolver stores 10-20x more data in RAM - turning complex data engineering projects into a few clicks.

Setup

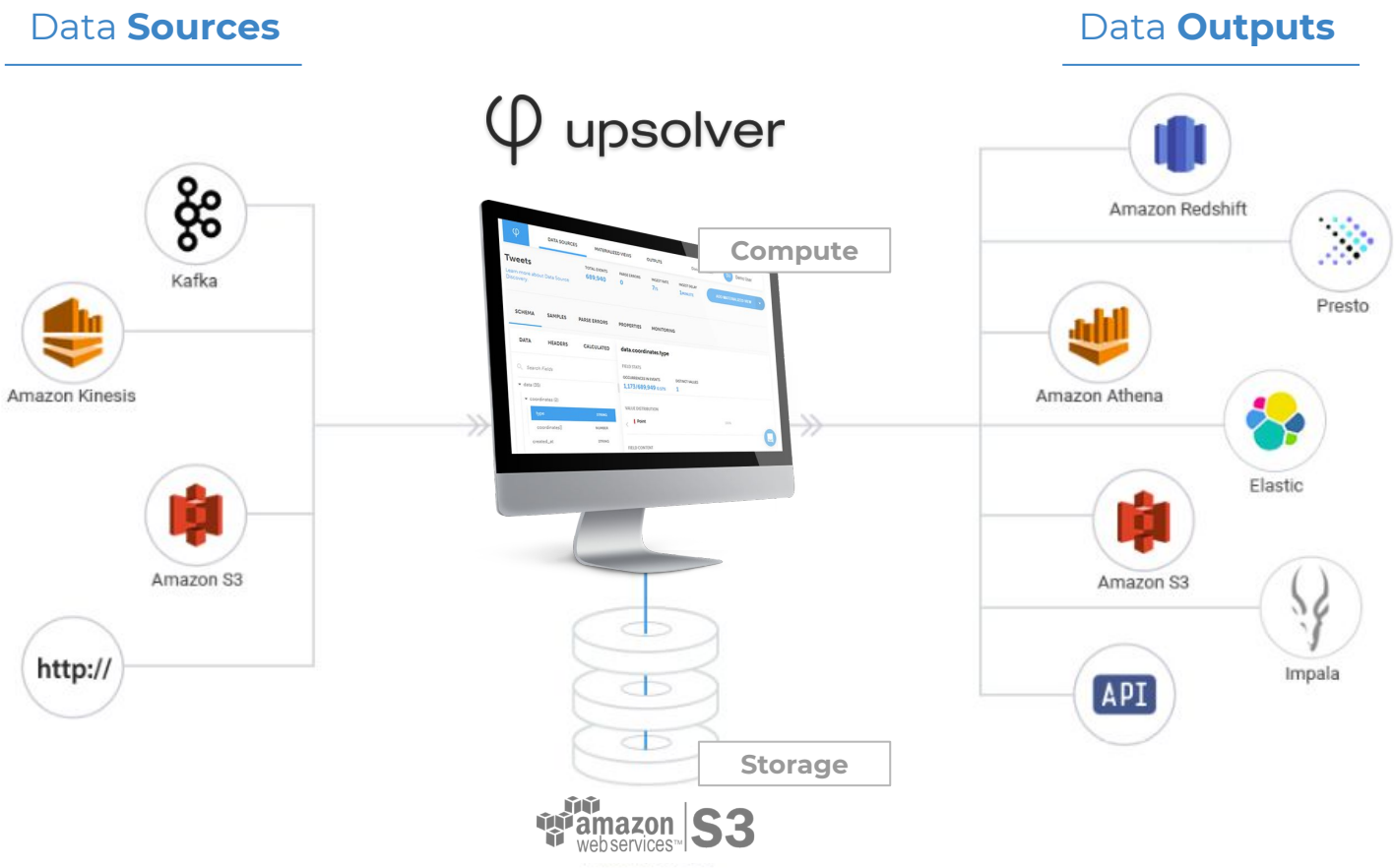
Deployment - Upsolver is deployed in the AWS cloud (Upsolver or customer account) and auto scales on-demand.

Configuration - Upsolver is easily configured using a friendly web application or a streaming SQL language instead of code.

Built-in Integrations & API

Deep AWS integration to Athena, S3, Kinesis, Redshift and Open source solutions like Presto and Impala and Kafka.
Upsolver also offers an API for real-time serving using HTTP REST.

Data Preparation **Done In Minutes**

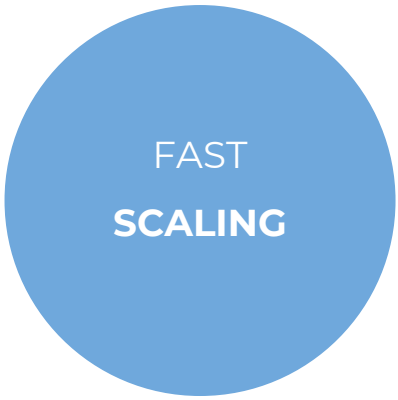


1 Generate More Insights

For streaming data, over 90% of the time spent generating insights is spent on data preparation by big data engineers.
Upsolver opens the big data engineering bottleneck with a self-serve data preparation for analysts, scientists, product managers and developers.

2 Scale Faster

With Upsolver, your analytics will easily scale from GBs to PBs. No storage or IT bottlenecks.



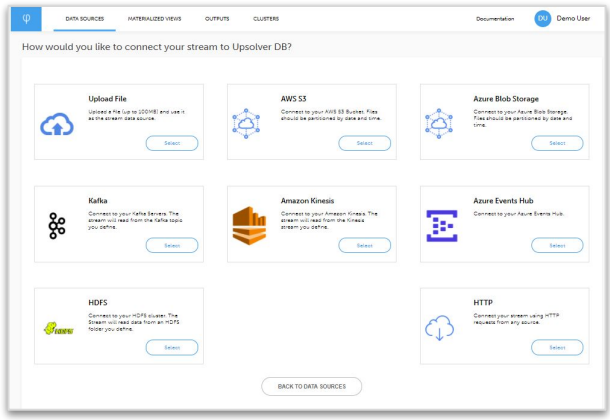
3 Reduce 50%-90% of your Analytics Cost

Upsolver offers unparalleled low TCO:

- Storage: Amazon S3
- Software: Pay per hour
- Hardware: Low cost servers (Spot)
- HR: No data engineering, no IT

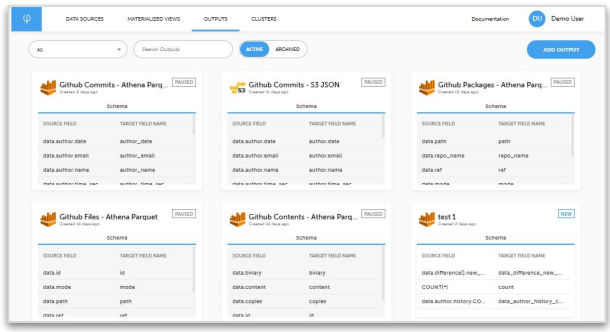
Reading Data

- Build-it integrations: S3 / Kafka / Kinesis
- Exactly-once read from Kafka / Kinesis
- Replay from S3



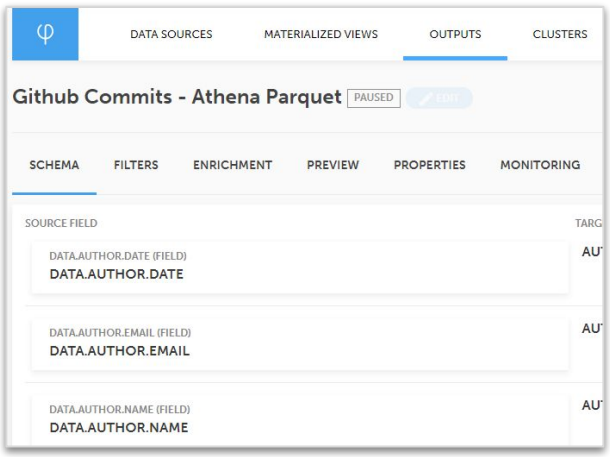
Delivering Data

- Built-in integrations: Athena / Redshift / S3 / Impala / Presto / ElasticSearch
- Manage partitions and files on S3
- Serving API for sub-millisecond queries



Enrichments

- 100s of transformation functions
- Extract geo data from IPs
- User agent parser
- User defined functions in Python
- Join with more data sources in ETL time

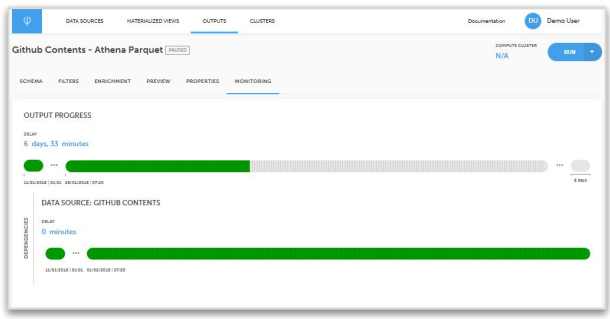


Aggregations

- Aggregated outputs (fact tables)
- Unlimited sliding time windows
- Session windows
- Accurate count distinct at scale
- Nested aggregations

Monitoring

- Built-in integrations to Customer Monitoring

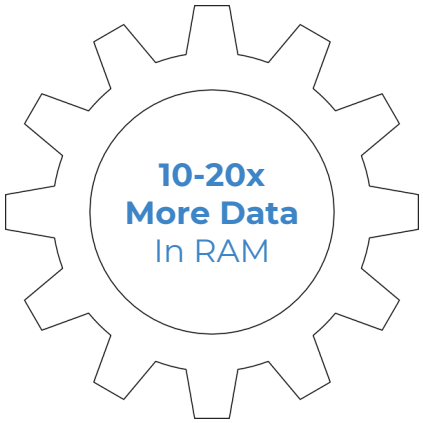


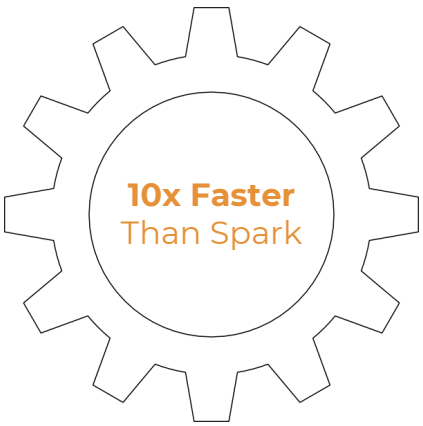
Powered By **Volcano™** Technology

In Memory Computing Running On **Jet Fuel**

Compression

- 90%-95% compression rate
- Read without opening the compression
- 10X compression speed acceleration



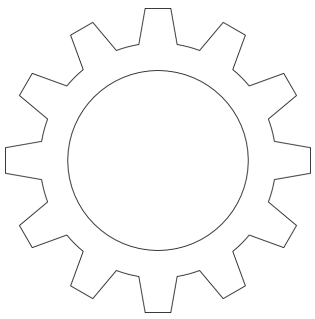


Performance

- 6x faster Avro/Parquet writers
- 1 millisecond queries from S3

Architecture

- In-memory processing over S3/HDFS
- Fault tolerant - no option of data loss
- Easily scalable using EC2 and S3



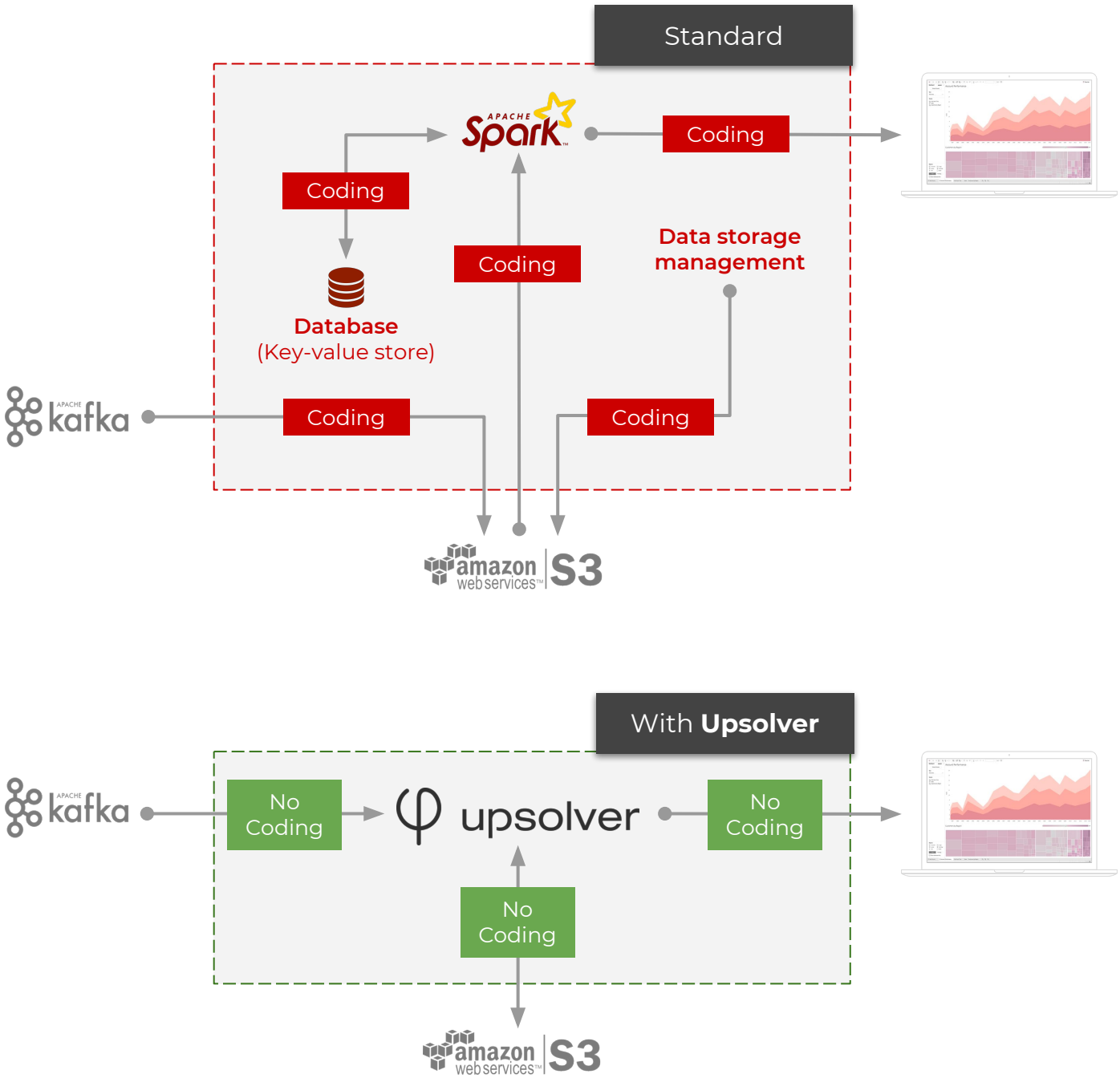
Streaming First

- Multiple streaming Aggregation functions
- Nested aggregations
- Unlimited sliding time windows
- Accurate count distinct at scale

Upsolver VS **Alternatives**

	Spark / hadoop	upsolver
IMPLEMENTATION TIME	Months	Hours
SKILLS REQUIRED	Code	Drag & Drop
INDEXING	Requires Additional Key-value Store	Included (Out-of-the-box)
DATA MODEL	RDDs	Hierarchical
MAINTENANCE	On-going (Both for cloud & on-prem)	Not Required
WINDOWING	Limited To Raw Data In-mem	Unlimited Using S3
SPEED	-	10x Faster

Reference **Architecture**



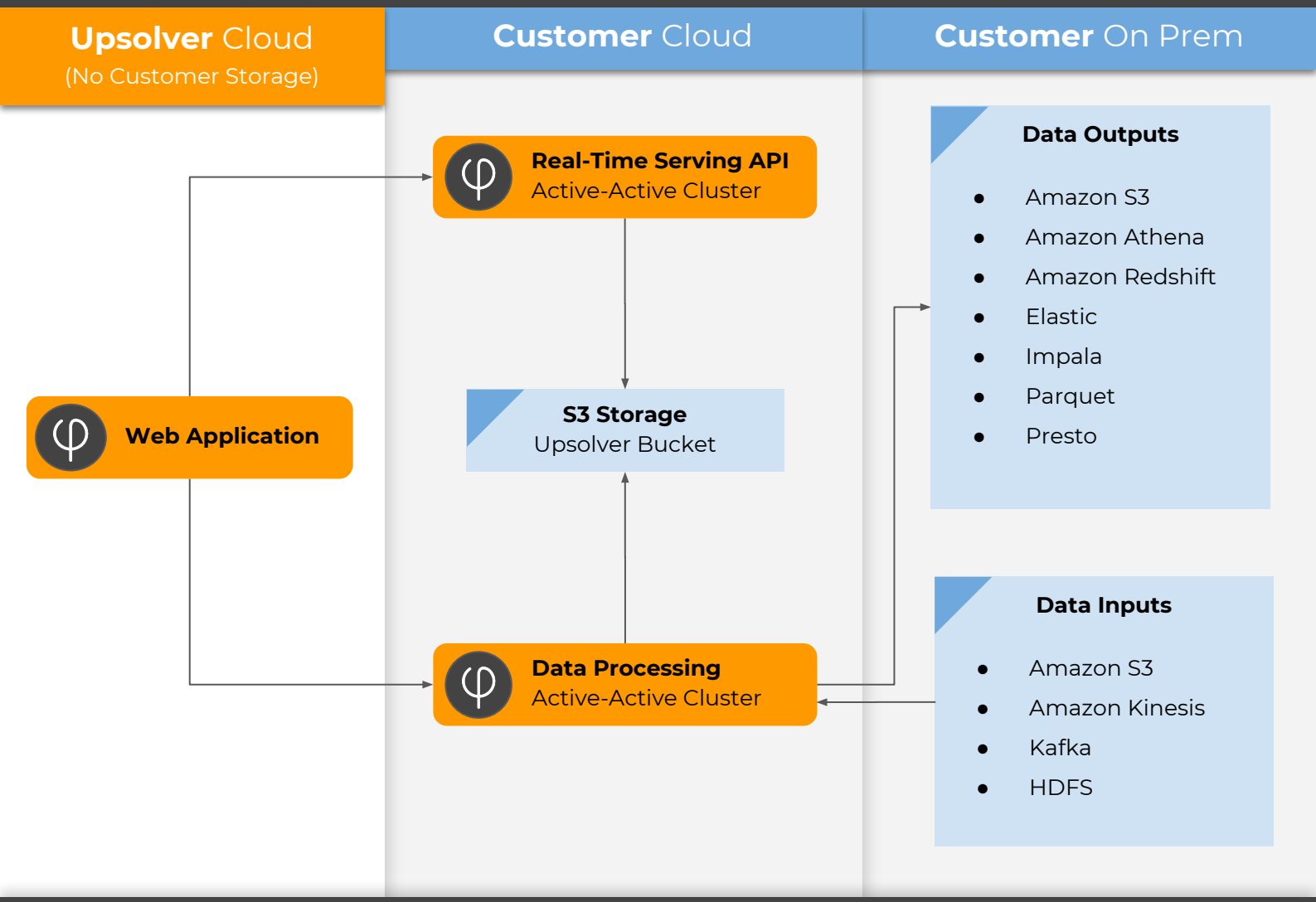
Without Upsolver

- Lengthy processes & pre-planning
- Multiple products and integrations
- ETL and data engineering are a bottleneck
- Replay historical data is very hard
- High storage costs in databases

With Upsolver

- No coding or ETL bottlenecks
- No IT overhead
- Low storage costs using S3
- One click to replay data from S3

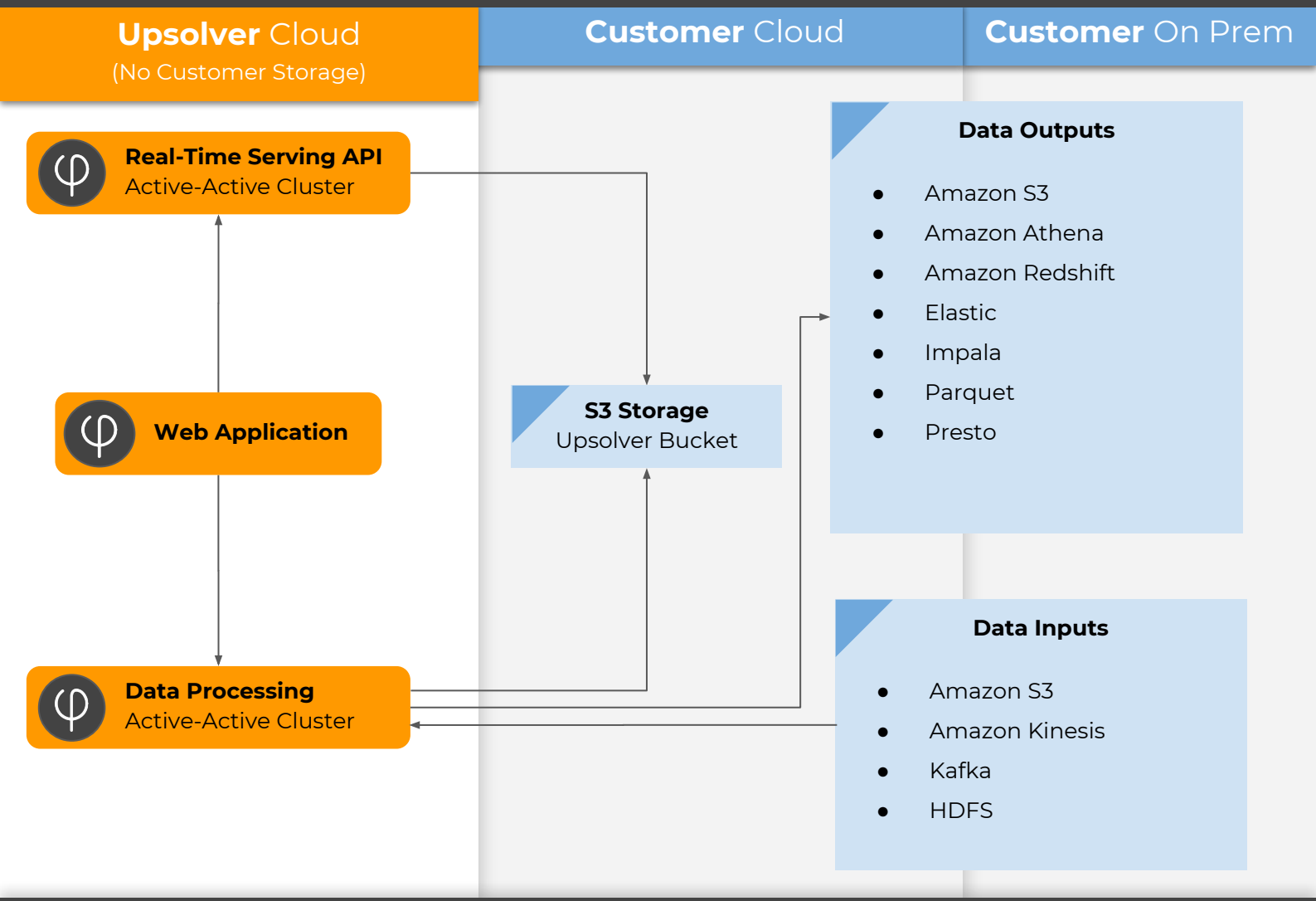
Private Cloud Deployment



Data Never Leaves **Your Cloud Account**

- **Storage** ONLY on Customer S3 Bucket
- **Spot instance** servers managed by Upsolver

Public Cloud Deployment



Compute In-memory
With Upsolver

Storage Only On S3

Your Data **Always** Remains **Private**

- **Storage** ONLY on Customer S3 Bucket
- **Processing** in-memory on Upsolver Servers
- **Serverless** / no IT

Implementation

Implementation **Example**

USE CASE

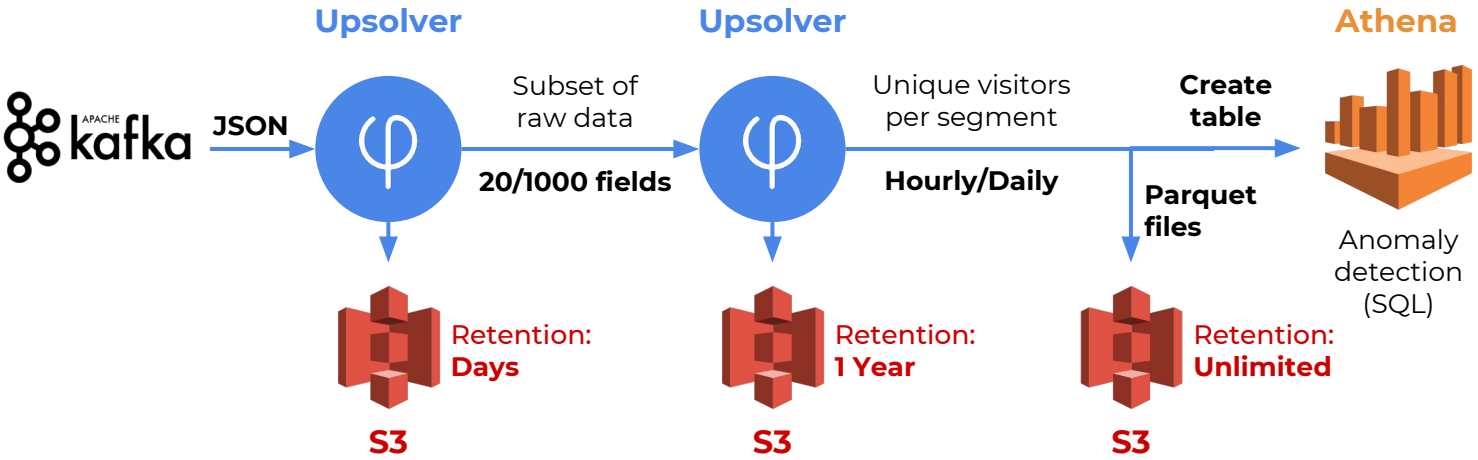
Marketing company, mobile SDK analytics

DATA

5GB/minute, JSON files, >1000 fields per JSON with hierarchies & arrays

RESULT

Implemented in an hour using self-serve, detected SDK anomalies



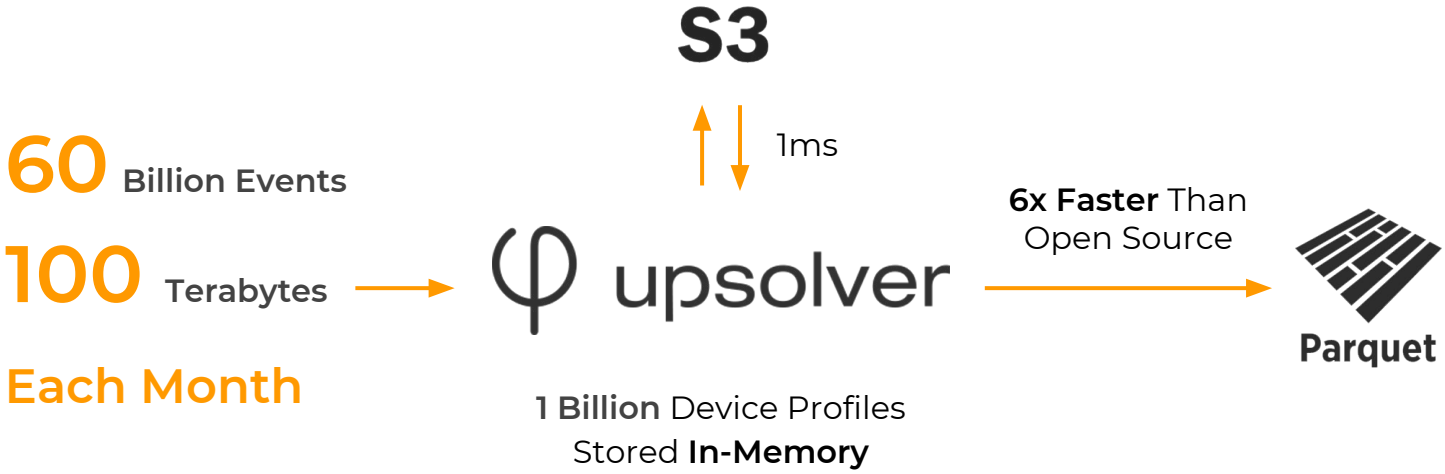
Implementation **Scale**

HARDWARE

4 Servers, 8 CPUs, 32GB RAM

IMPLEMENTATION TIME

1 HOUR



Permissions Required

- Upsolver managed bucket on Customer S3
- Read from Kafka / Kinesis / S3
- Write access to Athena / Redshift / etc
- Report metrics on Customer monitoring system

Operational Requirements

- **On Going Maintenance:** (2 hour / week)
  - Monitor Performance
  - Data Integrity
- **New ETLs:** (a few hours on-demand)

Upsolver Support (Optional)

- Dedicated account manager
- Email/Phone/Chat support
- 24/7

upsolver

In-Memory Data Preparation

- Schedule a Demo
- Start Free Trial