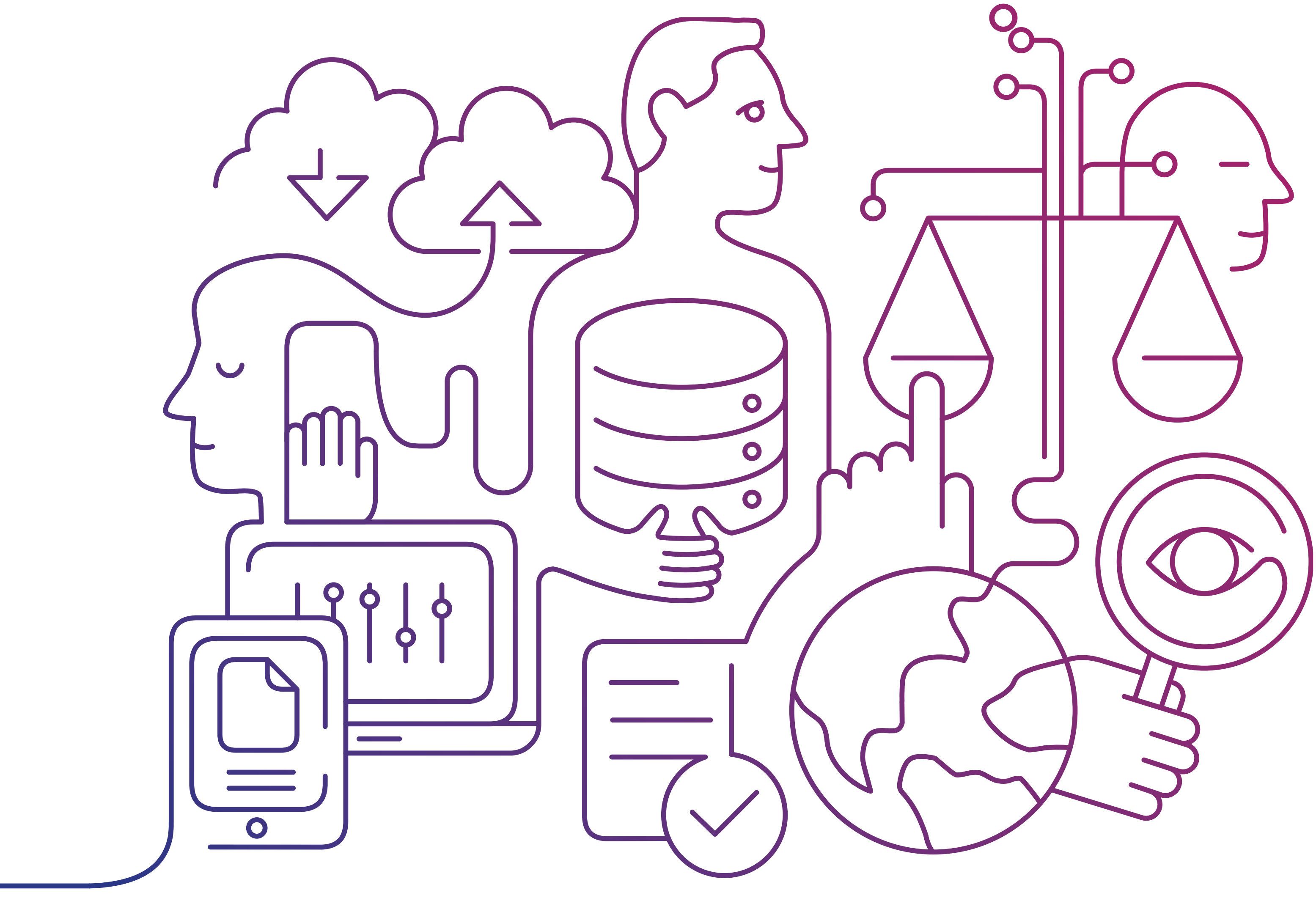
PRIVITAR

Scalability of the Privitar Data Privacy Platform

built on Modern Data Architectures



www.privitar.com

Using the Privitar Data Privacy Platform

Producing the best data-driven insights often involves the processing of very large amounts of safe, usable data.

Moreover, near real-time data processing is increasingly becoming a requirement as some contexts demand quick decisions.

Fortunately, modern data architectures for big data have developed at a rapid pace in the last few years and now provide infrastructure that supports the growth in dataset sizes. Tools such as Hadoop, Kafka, NiFi and, more recently, native cloud services have democratized data processing at the terabyte and even petabyte scale.

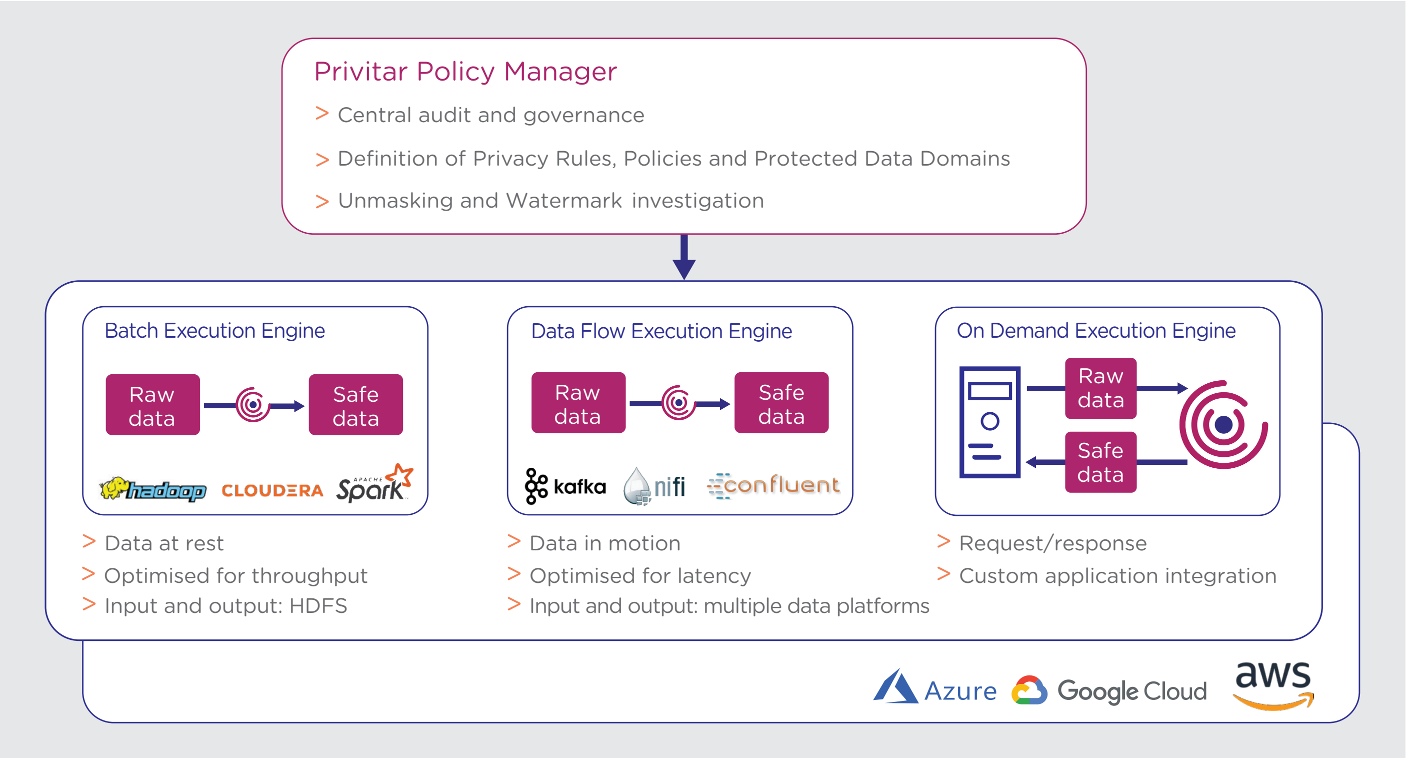
The Privitar Data Privacy Platform is architected and purpose-built for big data processing natively on these platforms to ensure that large scale is never an obstacle when protecting your customers' sensitive personal data.

Data Privacy Policies are defined centrally in the Privitar Policy Manager. Each Privacy Policy is packaged and sent to the enterprise data management platform of choice, where it is applied natively in the decentralized Execution Engines to produce a Privitar Protected Data Domain. By applying policies directly in the analytics environment where the data resides, Privitar exploits the scalable processing of these big data platforms.

The Batch Execution Engine is built using Spark and runs on Hadoop clusters, both on-premise and in the cloud. It is ideal when processing very large datasets at rest, or as part of an ETL pipeline. Thanks to the parallel data processing capabilities of these platforms, the Batch Execution Engine is the best choice to process large datasets in a data lake at high throughput.

The Data Flow Execution Engine is designed to process data in flight that is moving through your data pipeline. It is built on proven platforms, such as Apache Kafka, Confluent Platform and Apache NiFi.

The Privitar On Demand Execution Engine is perfect for an interactive style of data processing. It delivers Privacy Policies as a service that can be invoked via a simple HTTPs API call. As a stateless application it scales with the number of application instances across which the traffic is distributed.



Privitar Policy Manager

- Central audit and governance
- Definition of Privacy Rules, Policies and Protected Data Domains
- Unmasking and Watermark investigation

Batch Execution Engine

- Data at rest
- Optimised for throughput
- Input and output: HDFS

Data Flow Execution Engine

- Data in motion
- Optimised for latency
- Input and output: multiple data platforms

On Demand Execution Engine

- Request/response
- Custom application integration

Scalability and Token Vaults

Privitar's consistent tokenisation is a powerful tool that preserves the data linkability during the de-identification process. It works by building a secure "Token Vault" as a mapping between the raw values and the randomly generated tokens. Privitar supports the following Token Vault technologies:

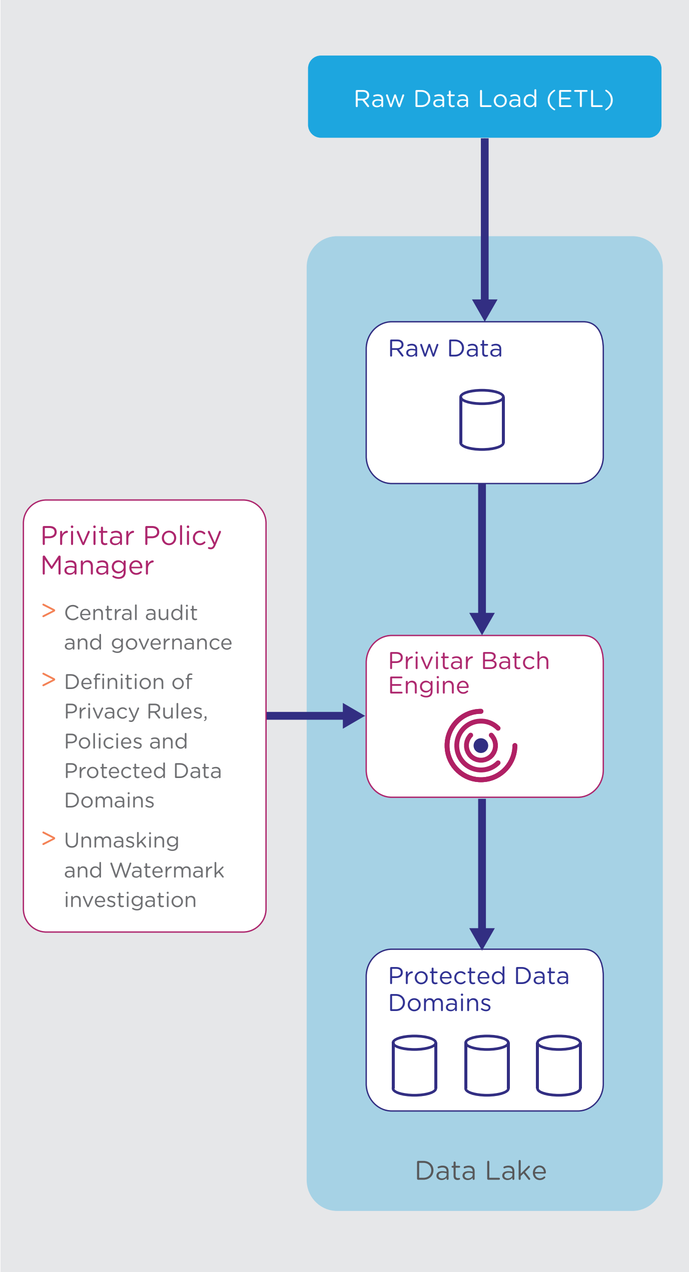
- HDFS (Batch only)
- HBase
- DynamoDB
- Oracle
- PostgreSQL

Different platforms offer different functional characteristics, but the choice of the correct Token Vault platform is a fundamental component for the resulting system performance. The Privitar Execution Engines are implemented on horizontally scaling platforms. Therefore, for consistent tokenisation-heavy workloads, it's important to select a Token Vault technology with sufficient IOPS to sustain the desired throughput or latency.

Token Vaults such as HDFS, HBase and DynamoDB offer simple horizontal scaling, while Oracle and PostgreSQL may be chosen for their widespread availability, support and vertical scaling capabilities.

Batch processing with Spark

The Batch Execution Engine allows Privitar to process very large datasets at high throughput. It uses Spark and Hadoop to parallelize the data processing and scale with the number of available nodes.


Raw Data Load (ETL)
Raw Data
Privitar Policy Manager
> Central audit and governance
> Definition of Privacy Rules, Policies and Protected Data Domains
> Unmasking and Watermark investigation
Privitar Batch Engine
Protected Data Domains
Data Lake

It should be used:

- To process batches of data at rest on existing data lake infrastructure
- If sensitive or PII data cannot leave the data lake where it is stored for any reason (policy or regulatory)
- If a dataset requires statistical privacy protections that consider the full dataset at once (autogen and k-anonymity)
- To unmask data or investigate watermarks within datasets that have been processed by Privitar

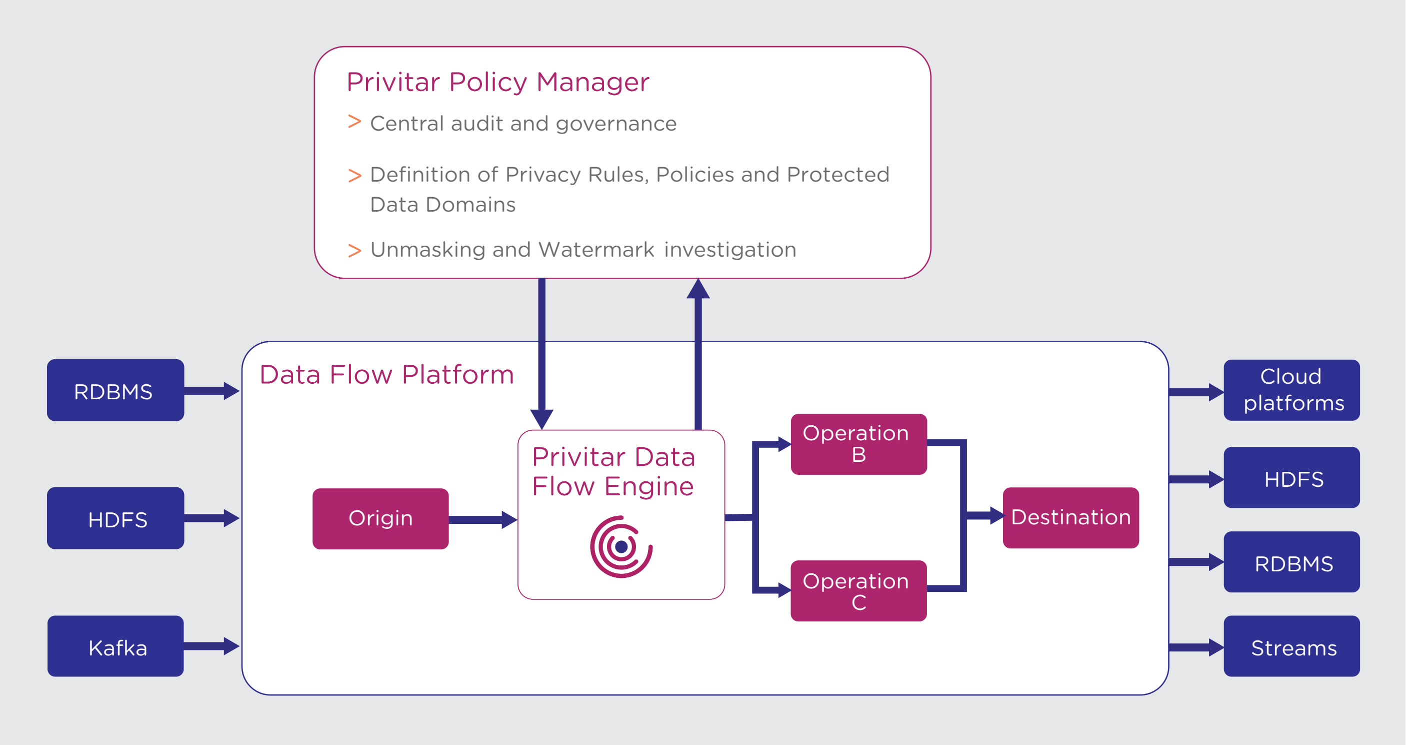
The Batch Execution Engine scales with the resources available in the data lake, such as the number of cores and memory available per executor.

Performance will depend on the specific characteristics of the infrastructure (such as available IOPS, clock speed, network bandwidth), but it can consistently achieve 10,000 to 15,000 de-identifications/second/core.

Customer example: batch processing at a health insurer

A Fortune 100 health insurer protected a dataset consisting of five years of member, claims and clinical data, for a total of 280 billion records, with one (1) trillion fields to be protected, for a total size of more than 35 TB.

The Privitar Batch Execution Engine applied the Privacy Policies on their Hadoop cluster and created a safe version of their dataset for use for analytics.


Privitar Policy Manager
> Central audit and governance
> Definition of Privacy Rules, Policies and Protected Data Domains
> Unmasking and Watermark investigation
RDBMS
HDFS
Kafka
Data Flow Platform
Origin
Privitar Data Flow Engine
Operation B
Operation C
Destination
Cloud platforms
HDFS
RDBMS
Streams

Stream processing with Privitar Data Flow

The Data Flow Execution Engine is used:

- For sensitive or PII data that must be protected as part of an ETL or processing pipeline and can't land on the destination environment (e.g., a cloud or a big data lake) in raw form for policy or regulatory reasons
- If data to be de-identified is already being processed on platforms such as Apache NiFi (Cloudera CDF / Hortonworks HDF) or Apache/Confluent Kafka
- To de-identify streaming data, where lower latency matters
- To transport data from one environment to another, taking advantage of an existing ecosystem of source and destination connectors

The Data Flow Execution Engine scales with the number of parallel pipelines or when running the pipelines in Cluster Mode.

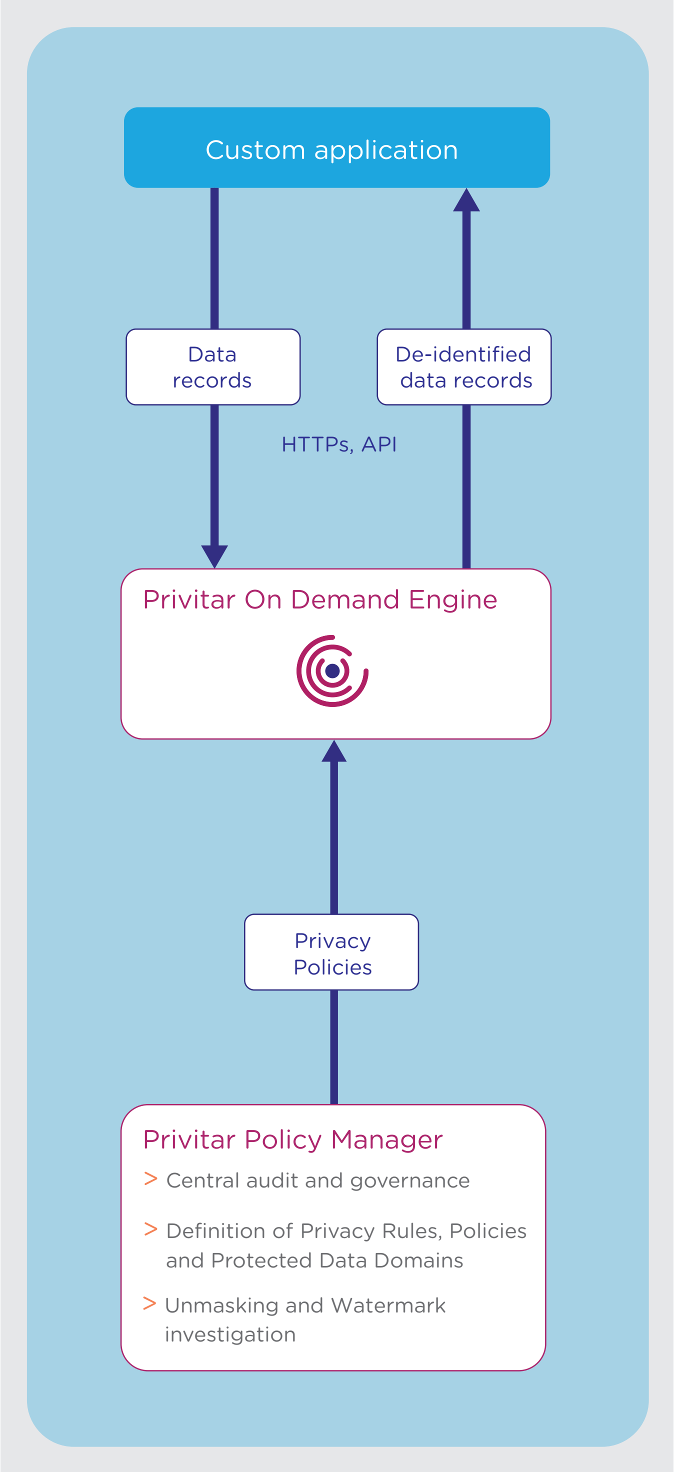
Performance grows with the IOPS for the Token Vault database. Different Token Vault technologies will offer different scalability options. For example, AWS DynamoDB offers the option to dynamically and automatically scale up or down with demand on the cloud. HBase scales by adding more resources to the cluster.

Customer example: NiFi data flow in a financial services company

A large financial services and insurance group is using the Privitar Data Flow Engine to de-identify data in a NiFi data ingestion pipeline.

They have achieved 140,000 tokenizations/second on their existing NiFi cluster, using an HBase token vault, which gives them horizontal scalability and the option of tokenizing consistently across both the Data Flow and the Batch Execution Engines.

Request/response processing with Privitar On Demand


Custom application
Data records
De-identified data records
HTTPs, API
Privitar On Demand Engine
Privacy Policies
Privitar Policy Manager
> Central audit and governance
> Definition of Privacy Rules, Policies and Protected Data Domains
> Unmasking and Watermark investigation

The On Demand Execution Engine should be used:

- If you need to apply the same Privacy Policies and protections used in Hadoop, Kafka and NiFi to custom applications or pipelines
- To integrate with a pipeline technology that is not natively supported by the Data Flow Execution Engine
- To de-identify smaller batches of records over HTTPs with a smaller latency than what would be possible in a batch job on Hadoop

The On Demand Execution Engine is stateless and can scale both vertically (by providing more cores to each instance) and horizontally (by deploying multiple instances of the service and distributing requests across them).

Performance is bound by the IOPS offered by the Token Vault database. Different Token Vault technologies will offer different scalability options. For example, AWS DynamoDB offers the option to dynamically scale up or down with demand. HBase will scale with adding more resources to the cluster.

Customer example: Privitar On Demand at a healthcare organization

As part of a deployment validation test for a healthcare customer, we have tested throughputs of Privitar On Demand up to 24,000 tokenisations / second / On Demand instance.

The dataset consisted of one (1) trillion data records in total, and was processed at a rate of 1.6M tokenizations/second using a cluster of 64 On Demand instances deployed in a scaled-up AWS environment backed by a DynamoDB token vault.

Scalability Summary

	Batch	Data Flow	On Demand
Ideal use cases	High throughput Existing data lakes Data at rest	Low latency (data streaming) Data in-motion	Custom application integration Request/ response pattern
Main scalability factors	Job executors and cores Available memory per executor Token vault IOPS for consistent tokenization	Data Flow platform cores Token vault IOPS for consistent tokenization	Total POD cores available Token vault IOPS for consistent tokenization
Horizontal scaling	Number of executors Horizontally scale the Token Vault	NiFi or Kafka Connect cluster size Horizontally scale the Token Vault	Load balance across multiple POD instances Horizontally scale the Token Vault
Vertical scaling	Yes	Yes	Yes
Example real world scale	1 trillion records de-identified on an existing Hadoop cluster	140,000 tokenizations / second with an existing Nifi cluster and HBase Token Vault	1.6 Million tokenizations / second with a POD cluster and a DynamoDB Token Vault
Token Vault options	HDFS HBase DynamoDB Oracle PostgreSQL	HBase DynamoDB Oracle PostgreSQL	DynamoDB HBase Oracle PostgreSQL

About Privitar

Organizations worldwide rely on Privitar to protect their customers' sensitive personal data and to deliver uncompromising data privacy that frees them to extract maximum value from the data they collect and manage.

With the powerful Privitar Data Privacy Platform, businesses can safely use data to gain valuable insights that support data driven decisions over intuition to innovate, identify market opportunities, accelerate time to market, acquire and retain customers, improve customer experience, and identify inefficiencies that ultimately grow revenues, reduce costs and increase profitability.

Founded in 2014, Privitar is headquartered in London and has offices in New York, Boston, Munich, Paris and Singapore.

Contact us:

e: info@privitar.com
t: +44 203 282 7136
w: www.privitar.com

PRIVITAR

@PrivitarGlobal

www.privitar.com