VMware

Data Solutions Overview

Mariia Bocharova
VMware Solution Engineer

June 2023

Data Transformation Services

Tanzu Labs Data Experts

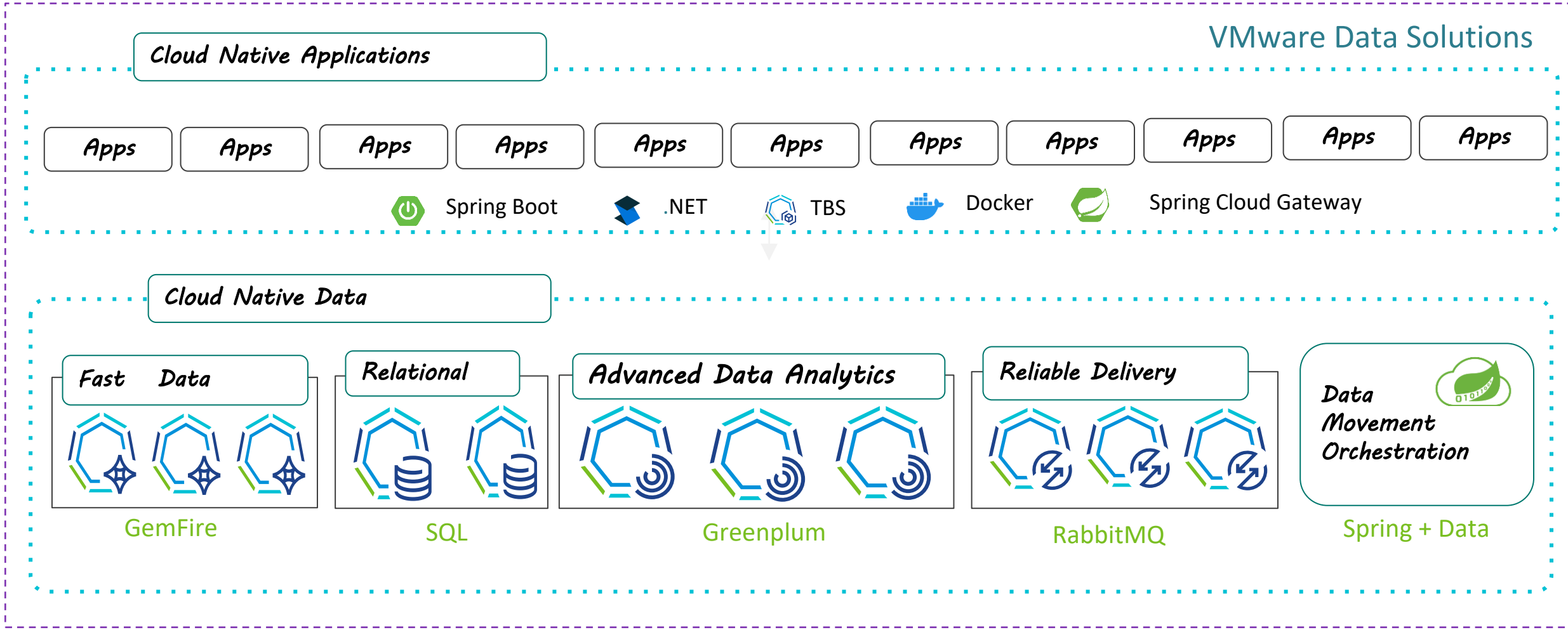
Enablement for your data modernization journey

- Enablement
  - “We don’t just drop in, deploy, and leave—we’ll work side by side with your team to build capabilities and instill a process that shows immediate and lasting impact. Empower your teams to deliver the products your customers need.”
- Modernize Data
  - We deliver Cloud Native based data applications using Kubernetes or VMs with reliability and high performance.
- Data Science
  - We can build Analytics, Artificial Intelligence and Machine Learning to drive value out of data.
- Data Movement
  - We develop data pipelines to reliably move data from one system to another in near real-time.
- Data Migration
  - We can help you permanently phase out legacy data systems for cost savings, improving efficiency and accessibility of your data.

VMware Data Solutions

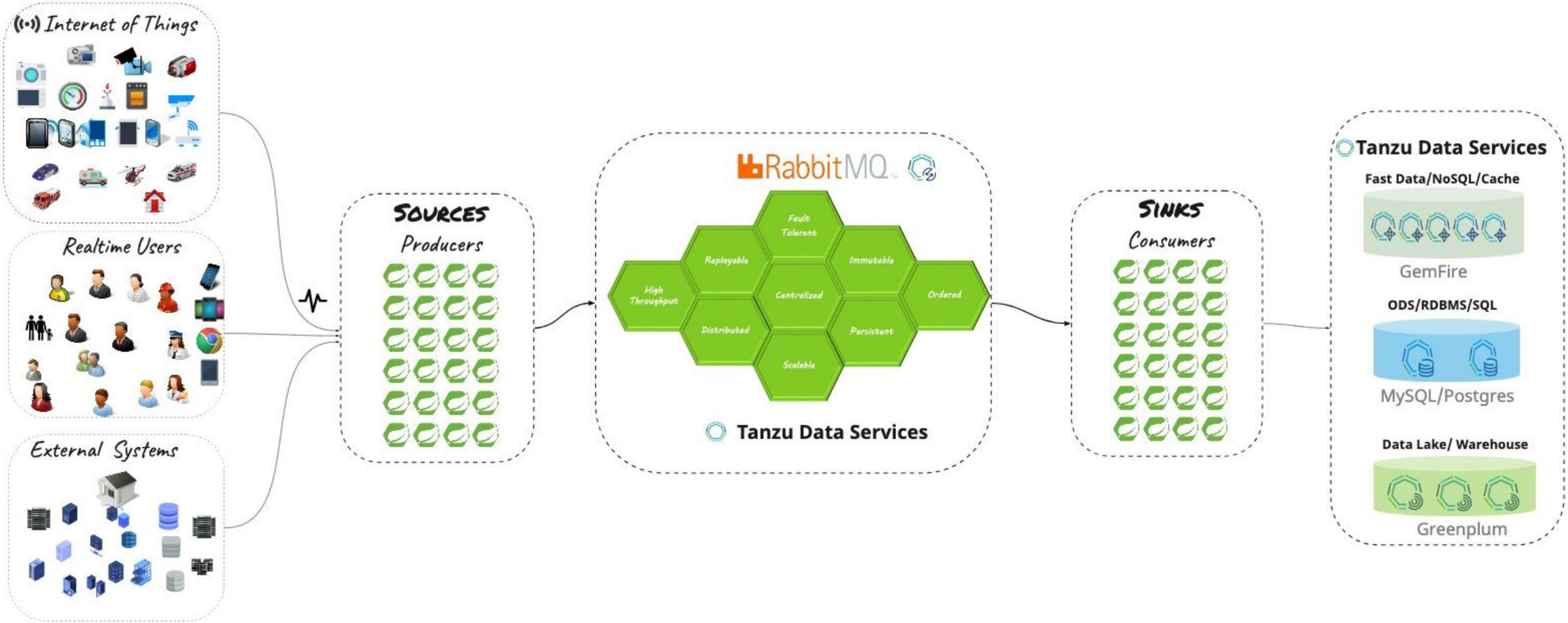
Information

Cloud Native – Architecture Patterns



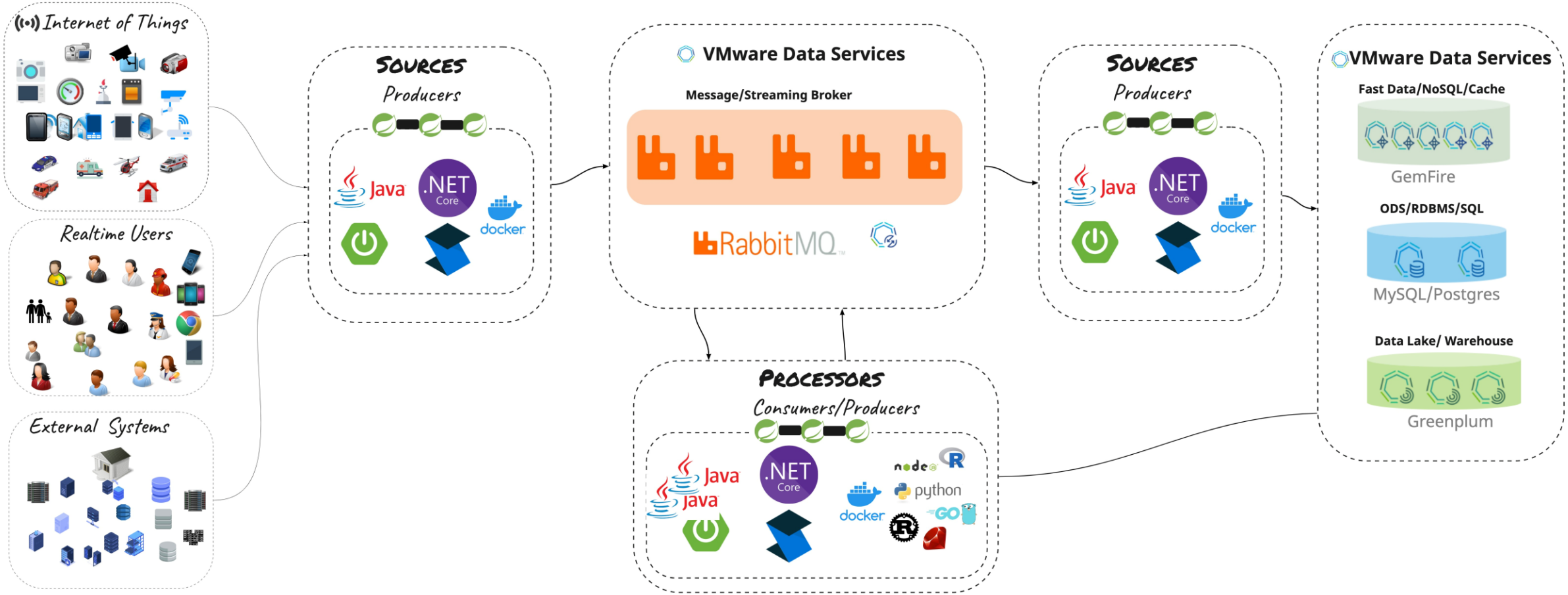
VMware Data Solutions

Modern Stack for Event Streaming Architectures



VMware Data Solutions

Modern Data Architecture



VMware Data Solutions Support

We are ready to partner with you when you need help

Full stack application and data support

- Support is provided by the Tanzu Support organization, available 24x7x365
- Support engineers are able to escalate to product engineers
- Further, App Modernization Technical Account Managers are available and can provide
  - Patching and upgrade planning
  - Weekly Care Log reports
  - Interviews or discussion session with Product Management
  - Root Cause Analysis Reports
  - Remote assistance
  - Product feature requests
  - And more

Target Response Times	Business Hours of Availability
Critical (Severity 1)	30 minutes or less (24x7)
Major (Severity 2)	2 business hours (12x5)
Minor (Severity 3)	8 business hours (12x5)
Cosmetic (Severity 4)	1 business day (12x5)

VMware Data Solutions

Cloud ready backing-services, Self Provisioning, On-Premise and Multi-Cloud, Scalable, HA - Fault Tolerant and secured access

Based on open source
24/7 Support

VMware SQL

Relational MySQL, Postgres or SQL Server databases

VMware GemFire

Fast Data NoSQL and Cache support (Redis compatible)

VMware RabbitMQ

High throughput broker for reliable messaging delivery

VMware Greenplum

Data Lake/Warehouse
OLAP via OLTP Consolidation

Data Management

VMware Data Solutions

Infrastructure for running modern apps and backing services with consistent, conformant Kubernetes everywhere.

Data Management
Management for VMware Data Solutions instances

I need a fast data store

GemFire
Fast In-Memory data store for Caching (Redis compatible, Transactional and NoSQL

I need to replatform a relational database

SQL
Relational MySQL or Postgres database for Transactional or Analytic data processing

drive analytic value of out tons of existing data

Greenplum
Massively Parallel Processing (MPP) Postgres for Big Data store for analytics, Machine Learning and Artificial Intelligence

I need reliable messaging delivery

Rabbit MQ
High throughput broker for reliable messaging delivery

I need flexible and manageable data integrations

Spring + Steeltoe + Data
Data services, connectors and integration orchestration for data pipelines (ex: ETL, streaming, etc.)

Features

- ✓ Cloud deployed backing-services
- ✓ Self Provisioning
- ✓ On-Premise and Multi-Cloud
- ✓ Scaling
- ✓ HA - Fault Tolerant
- ✓ Secured access
- ✓ Based on open source
- ✓ 24/7 Support

vmware

RabbitMQ

I need reliable messaging delivery

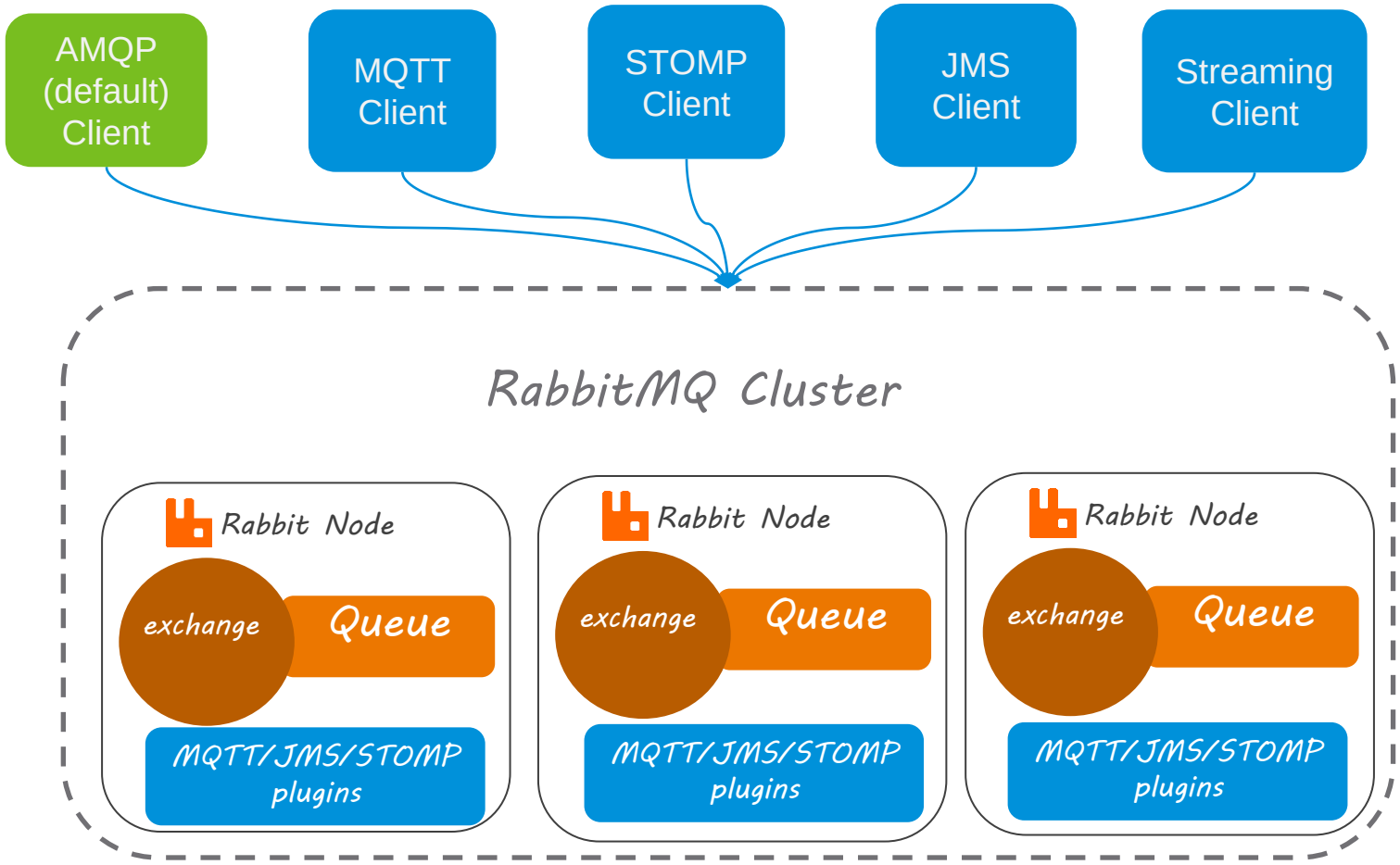
Message Broker

Reliable Messaging

RabbitMQ Protocols

RabbitMQ supports various interfaces using plugins that ships in the core distribution.

- Supports TLS connections
- User authentication/ authorization
- Exchanges used a Publishing destination abstraction
- Store messages in Classic Queues, Quorum Queues or Streams

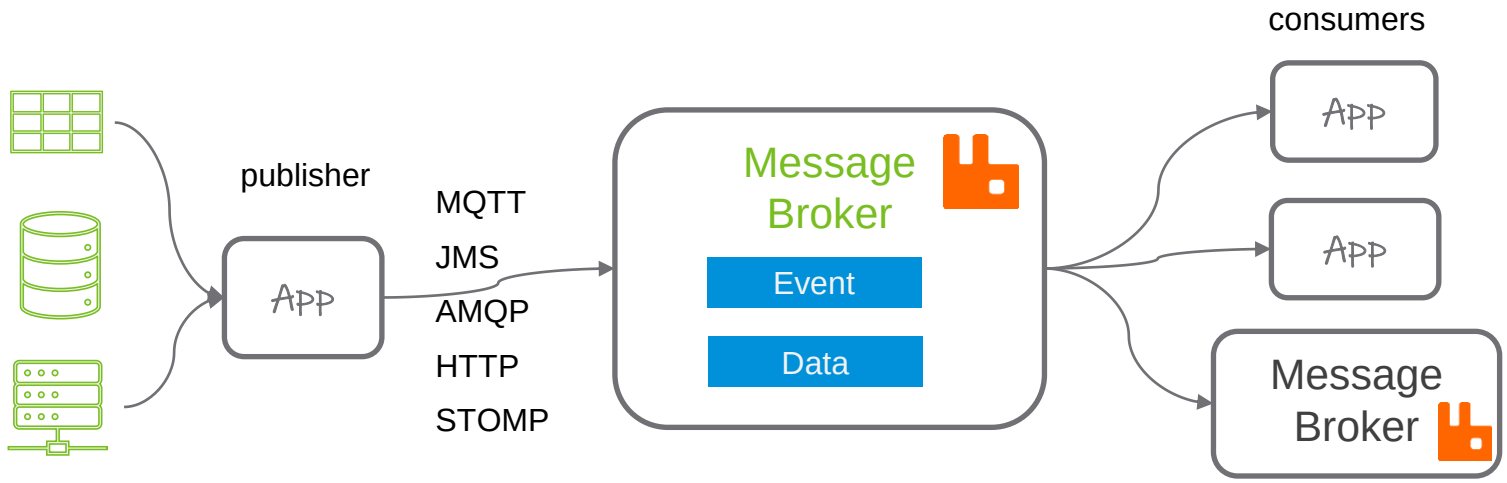


RabbitMQ

Best of Streaming & Messaging

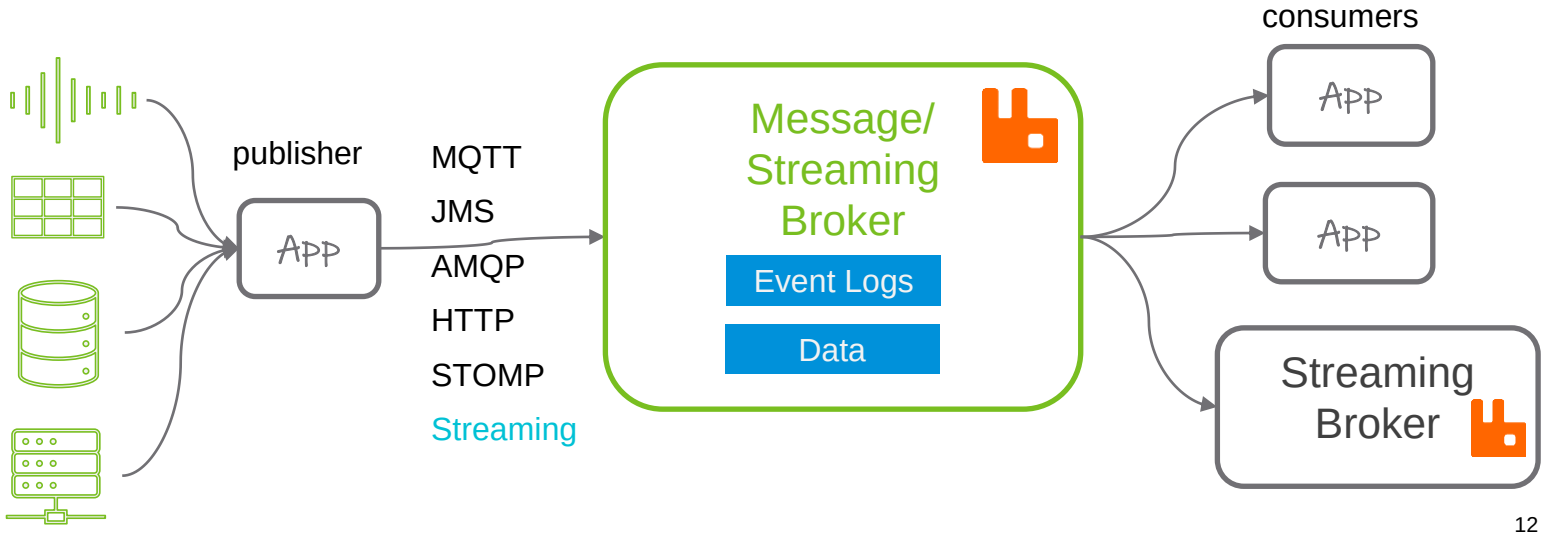
Message Broker

- Message removed after reliably delivery
- Low latency
- Flexible routing within broker



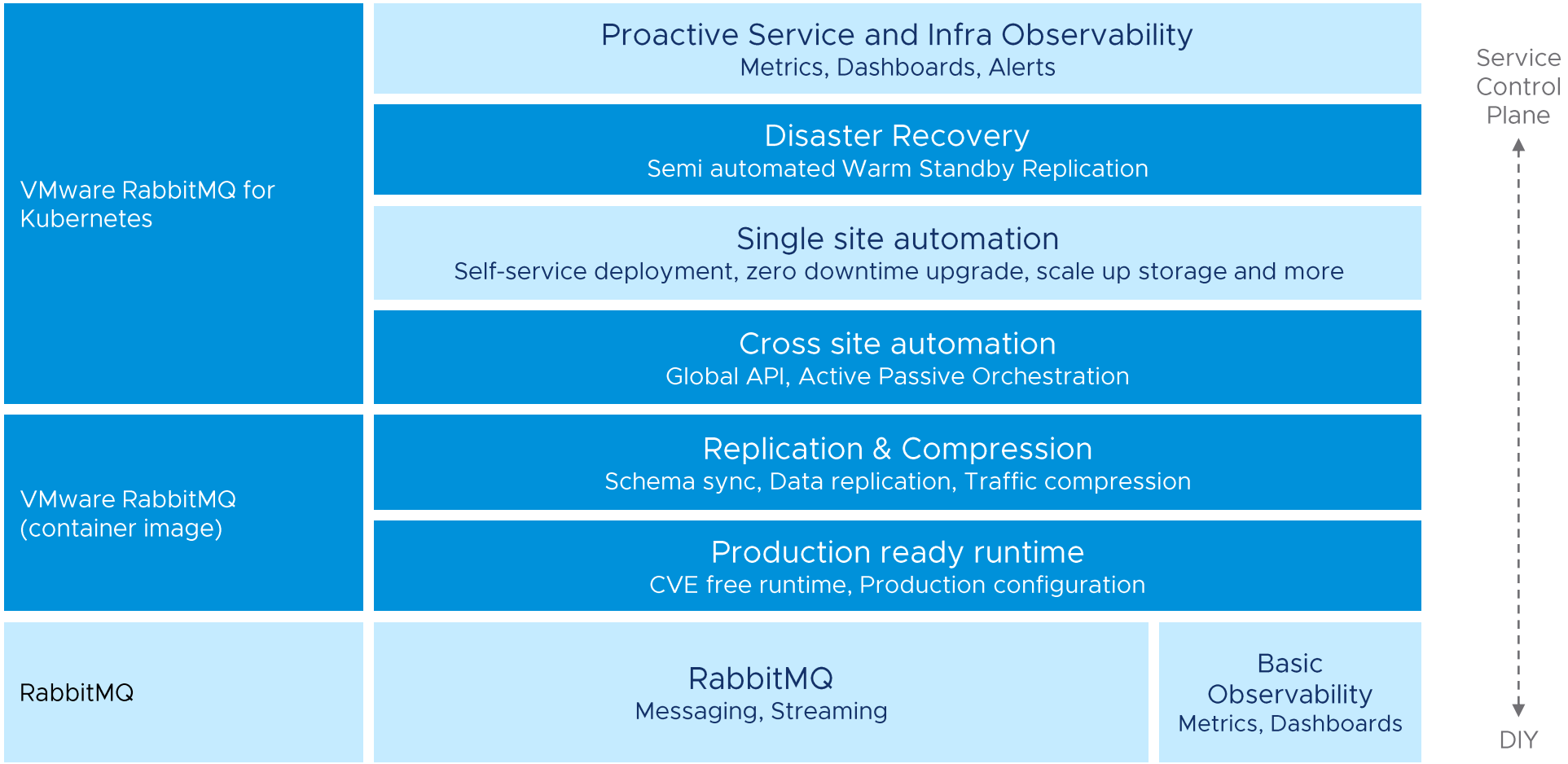
Streaming Broker

- Replay messages after reliably delivery
- High throughput
- Routing within App



VMware RabbitMQ

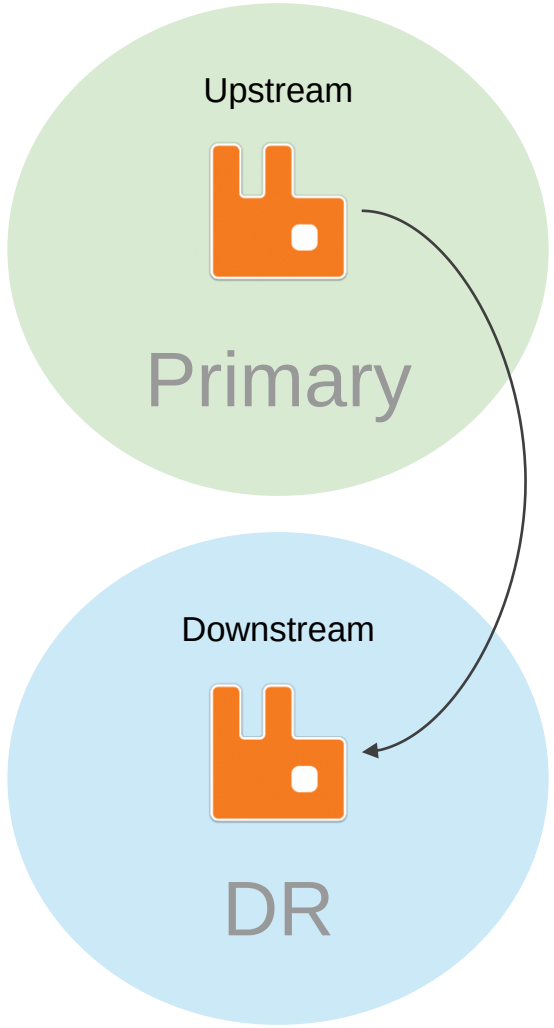
We develop these offerings for our customers and users



VMware RabbitMQ: Schema/Warm Standby replication (Commercial)

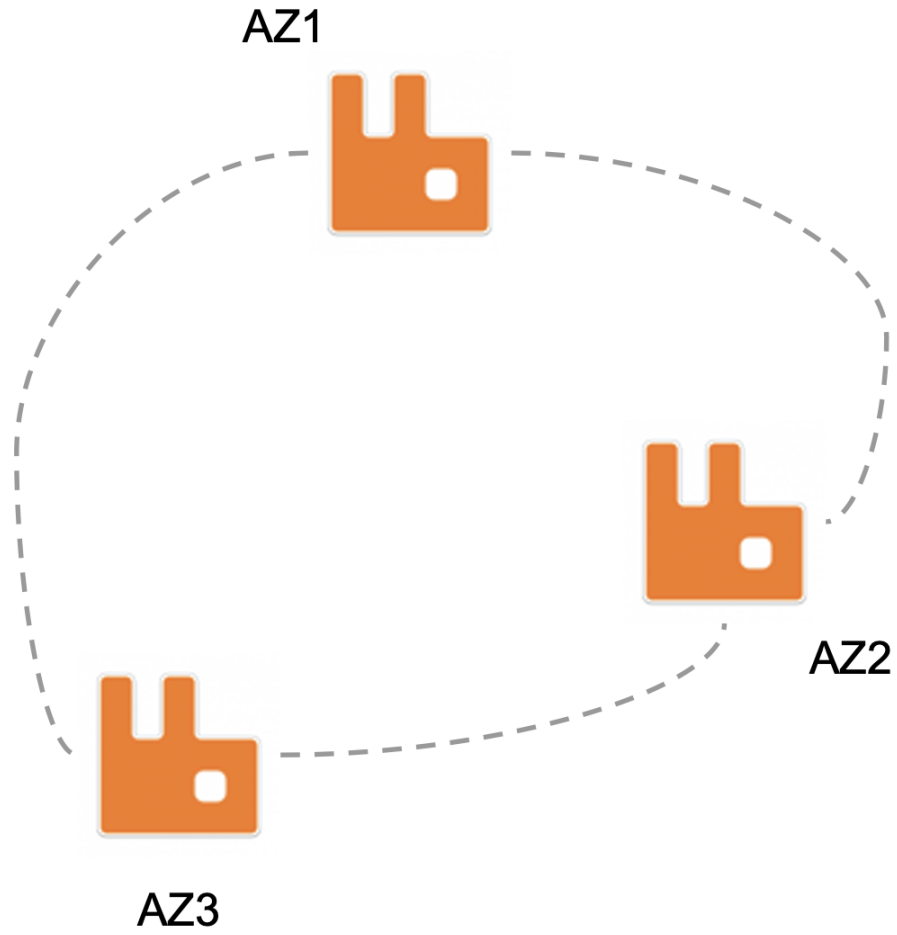
Enterprise-grade disaster recovery

- **Fast, data-safe message replication**
  - Uses the latest in RabbitMQ protocols and best practices
- **Automatic downstream cluster protection**
  - Automatically prunes messages that have been processed on the active upstream site
- **Easy setup**
  - No need to calculate or assume message throughput rates to configure message expiration/TTL (time to live)
- **Provides faster failover**
  - Downstream applications will only see messages that haven't been processed on the primary site, reducing the time to recovery



VMware RabbitMQ: Compression (Commercial)

Reduce IaaS costs in your high availability cloud deployments



- When would you use this plugin?
  - High availability replicated queues across Azs
- What benefits would you expect to see?
  - IaaS cost savings
- How does it work?
  - RabbitMQ nodes communicate with their peers and CLI tools using dedicated TCP connections, optionally protected with TLS.
  - https://www.rabbitmq.com/clustering-compression.html
  - Uses zstd by default, which has the best CPU to latency & compression ratio balance

VMware RabbitMQ: HashiCorp Vault Integration (Commercial)

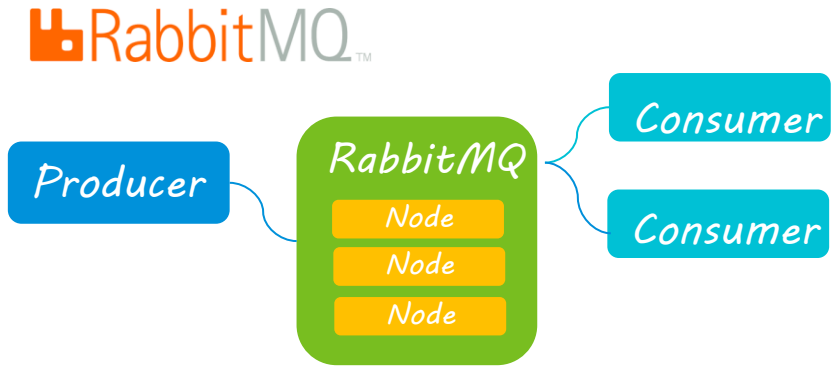
Enterprise-grade disaster recovery

- **Supports HashiCorp Vault secret management on Kubernetes**
- **Store RabbitMQ cluster user credentials**
- **Manage TLS server certificates**
- **Strengthens stored secret encryption at rest**
- **Automated rotations of passwords TLS certificates**

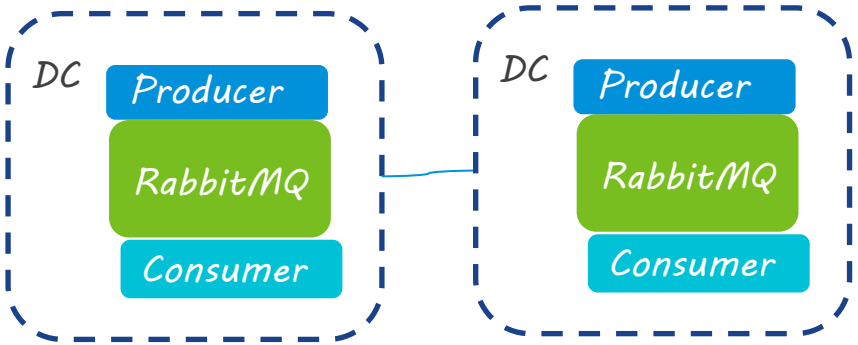
RabbitMQ Use Cases

Example Requirements

I need reliable messaging delivery



- As a user I need **FAST** Low latency processing
  - In memory messaging support
  - Push based message delivery by default
  - Scale out architecture
  - Time sensitive messages with strict Service Level Agreements (SLA)
- As a user I need **high throughput** event log processing
  - In the range of a million messages per second
  - Time based message replay capability
- As a developer I need a simple **user-friendly** implementation
  - Most popular open-source message broker
  - Develop cross-language messaging with your favorite programming languages, like Java, .NET, PHP, Python, JavaScript, Ruby, Go, and more.
  - Back pressure flow control support
  - Built-in management dashboard user interface
  - Out of box monitoring and alarms
- As a user developer, I need flexible routing
  - Abstraction between producer exchanges and one more consumer's queues
  - Message routing based on application provided keys (routing keys) and or message headers patterns
  - Distributed local data center architecture or cross data center Wide Area Network replication



RabbitMQ Observability

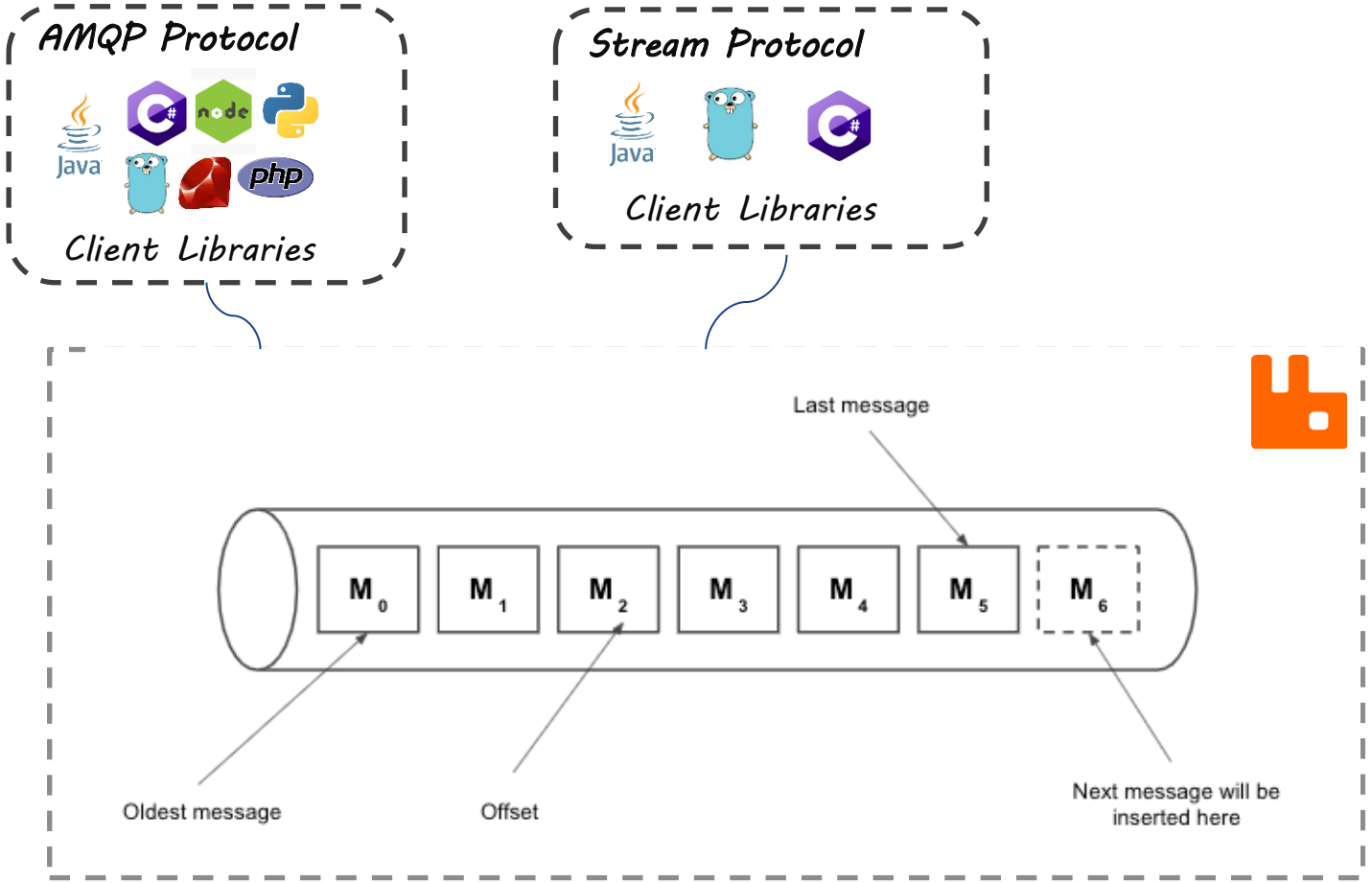
Monitoring and Alerting

- Collecting, processing, aggregating, and sending metrics to a prometheus compatible platform.
- Visualize key point indicators with out of box dashboards.
- Out of box dashboards with Tanzu Observability (Aria) and Grafana/ Prometheus.
- Allows configurable actionable alerts based on best practices and anti-patterns

Messaging Streaming

High Throughput Replay Messages

- Kafka like event logging
- **Large fan-outs:** when several consumer applications need to read the same messages.
- **Replay / Time-traveling:** when consumer applications need to read the whole history of data or from a given point in a stream.
- **Throughput performance:** when higher throughput than with other protocols (AMQP, STOMP, MQTT) is required.
- **Large logs:** when large amount of data need to be stored, with minimal in-memory overhead.




```
channel.queueDeclare(
  "my-stream",
  true,          // durable
  false, false, // not exclusive, not auto-delete
  Collections.singletonMap("x-queue-type", "stream")
);
```

```
channel.basicConsume(
  "my-stream",
  false, // not auto-ack
  Collections.singletonMap("x-stream-offset", 0),
  (s, delivery) -> { }, // delivery callback
  s -> { } // cancel callback
);
```

GemFire

I need a fast data store?

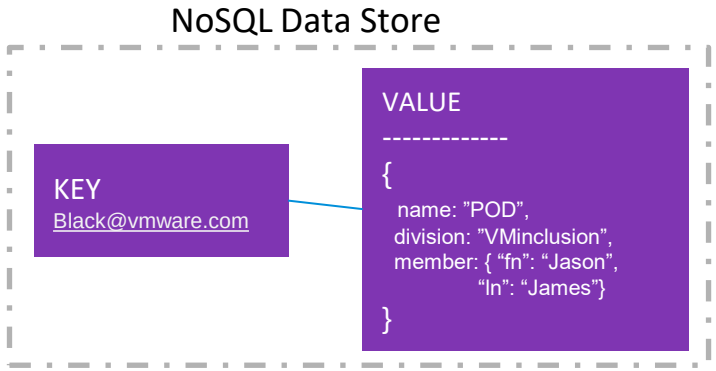
Powered by
Apache Geode

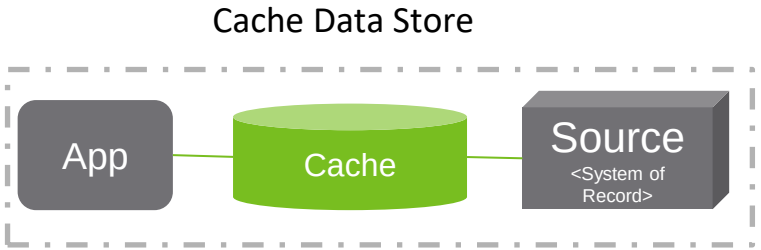
GemFire

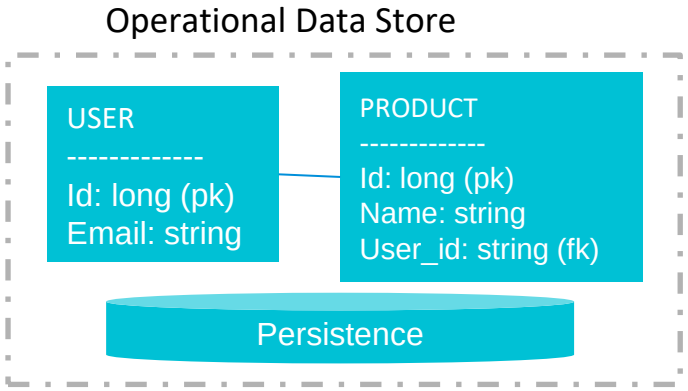
Use Cases

I need a fast data store?

- **NO SQL** data store
  - Fast lookup by key identifiers In-Memory
  - Query and full-text search access
  - Horizontal scalability support
  - High-Availability & Fault Tolerance support
  - Triggers/Event notations (client/service listeners)
  - Distributed In-Memory processing (ex: stored procedure, Map-Reduce, functions)
  - WAN replication (Active-Active or Active-Passive)
- **Cache** data store
  - Compatible with Redis applications clients
  - API exposed to user interfaces with a real-time interface
  - < 1 second response times
  - Expire cached entries as needed
- Transactional **Operational** data store
  - Persistent with STRONG Consistency – ACID compliance
  - System of record







GemFire

Fundamentals

Core components

- Data Node – In-memory data storage (a.k.a. Cache Server)
- Locator – clients and data nodes controller

Add Data Nodes as needed

- Handle data growth
- Increased processing demands of clients
- Supports resiliency

GemFire cluster

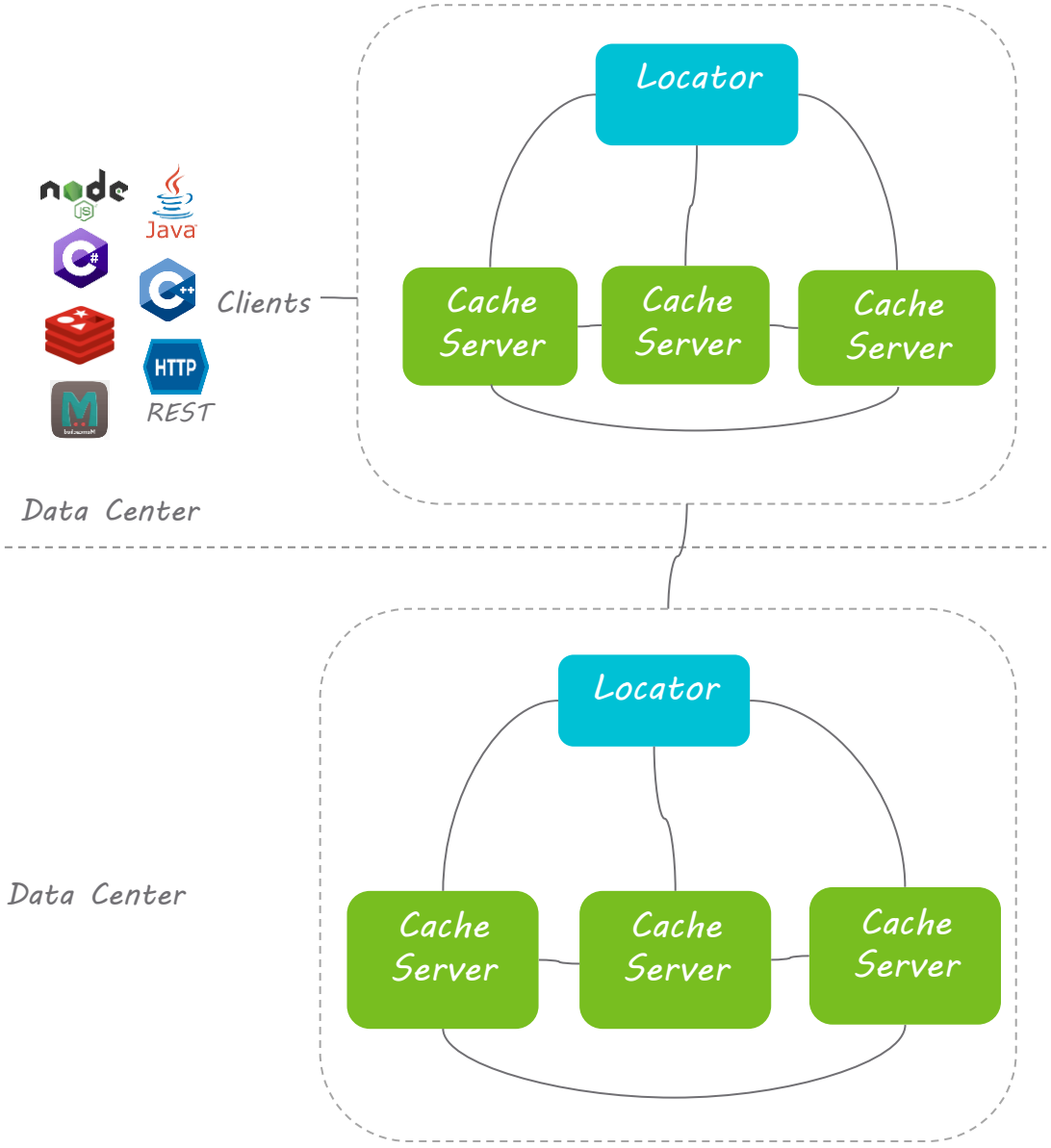
- Connected locators and data nodes

Clients

- Various supported client libraries

WAN Replication

- Replication data across data centers for disaster recovery (DR)
- Active-Active or Active-Passive

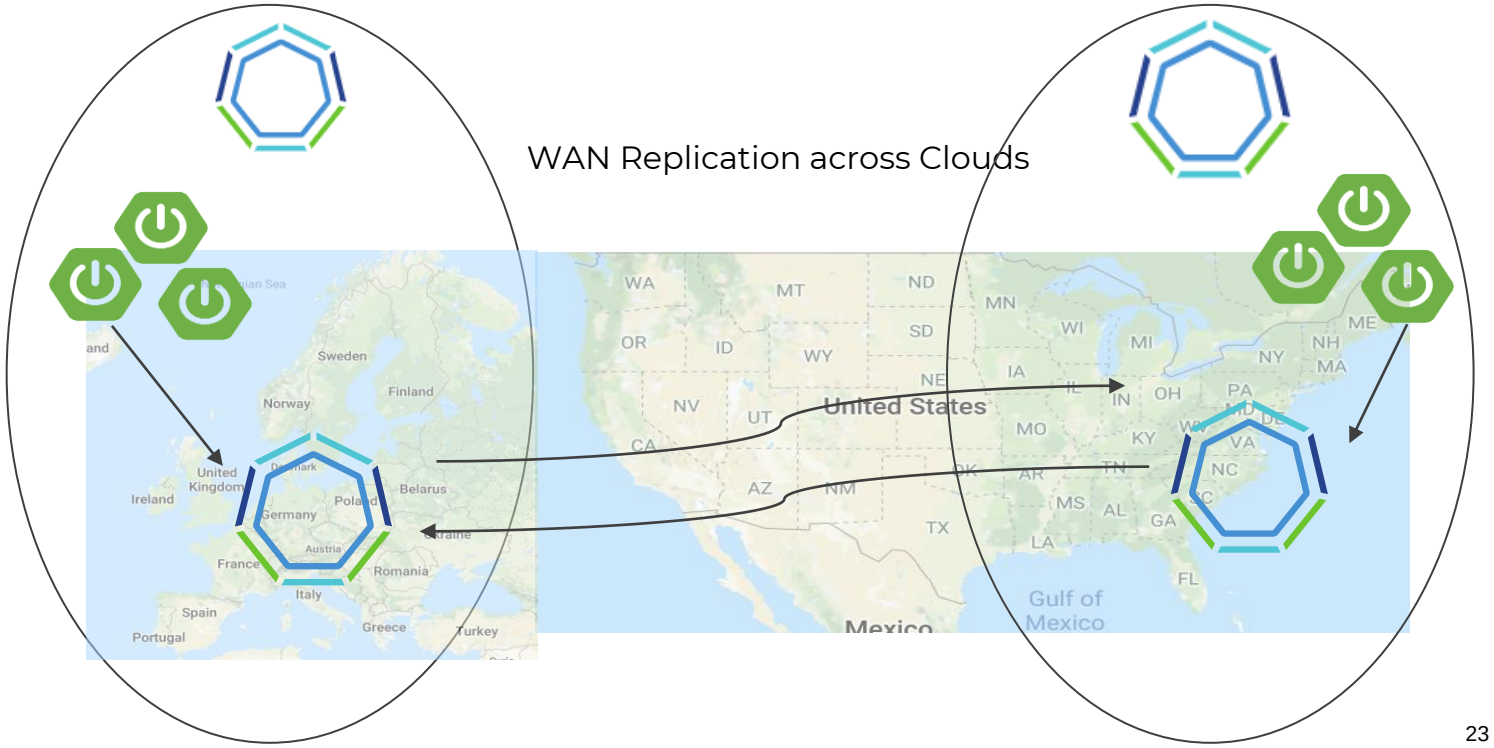


VMware GemFire

GemFire as a Service

- Multiple Cloud Ready
- Easy provisioning of environment instances
  - cf create-service p-cloudcache dev-plan app-cache -c '{"num_servers": 5}'
  - kubectl apply –f clusters-5-nodes.yml
- Elastic scale-out
- Self Healing
- High availability and Fault Tolerance
- Multi-site replication

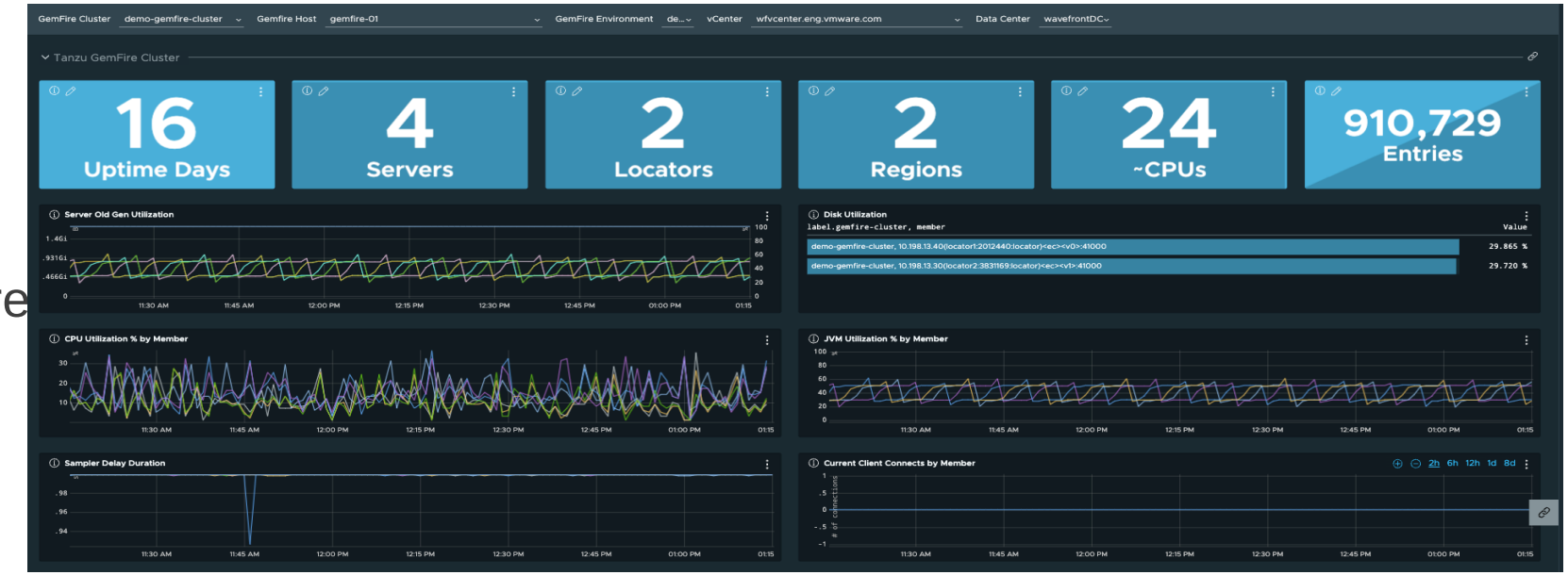
```
aversion: gemfire.tanzu.vmware.com/v1
kind: GemFireCluster
metadata:
  name: gemfire1
spec:
  image: registry.pivotal.io/tanzu-gemfire-for-kubernetes/gemfire-k8s:1.0.0
  locators:
    replicas: 2
  servers:
    replicas: 3
```



VMware GemFire

Observability

- VMware GemFire integration with VMware Observability
- Key Point Indicators
- Memory, CPU, Disk Storages and Network utilization
- Full Stack Dashboards on vSphere

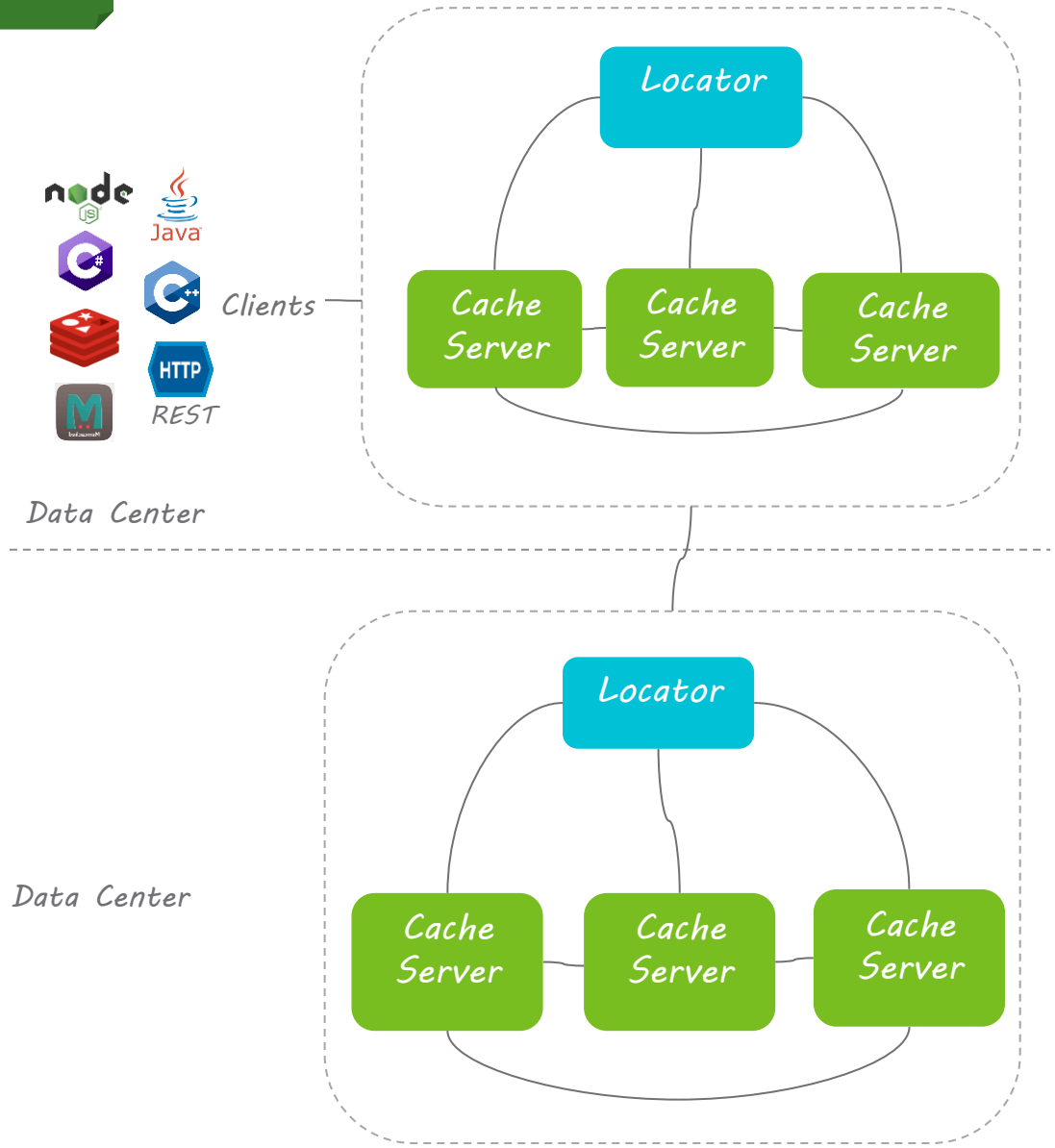


GemFire

Overview

I need a fast data store?

- **NO SQL** data store
  - Fast lookup by key identifiers In-Memory
  - Query and text-search access
  - Horizontal scalability support
  - High-Availability & Fault Tolerance support
  - WAN replication
  - Triggers/Event notations
  - Stored procedure data processing need
- **Cache** data store
  - API exposed to user interfaces with a real-time interface
  - < 1 second response times
  - Expire cached entries as needed
- Transactional **Operational** data store
  - Persistent with STRONG Consistency – ACID compliance
  - System of record



I need to replatform a relational database?

VMware SQL

MySQL & Postgres

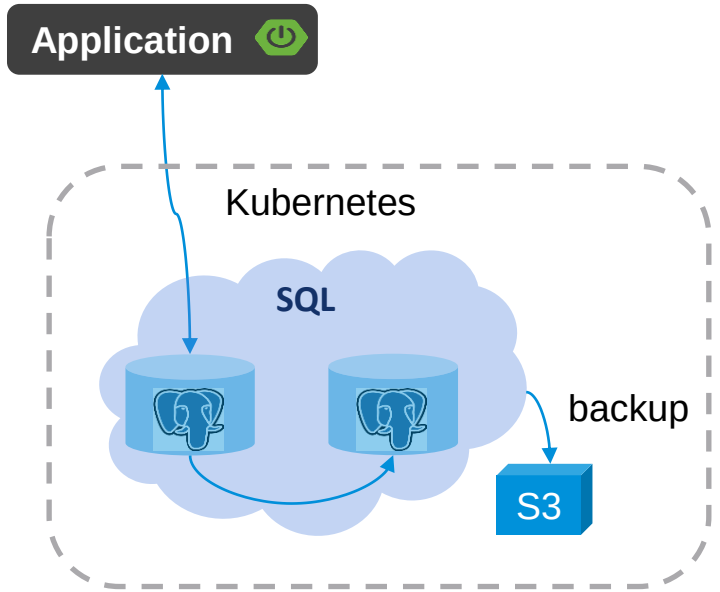
VMware SQL

Use Cases

- **Relational** database for transactions system, data marts or data warehouses
  - **MYSQL & Postgres** used for real time Online Transactional Processing (OLTP) data
  - **Postgres** also used for Online Analytical Processing (OLAP)

- **PaaS Based Operators**
  - Simplified provisioning and maintenance
  - On-premise and cloud-based deployments
  - Self-service
  - Blue – Green deployments
  - Smoothly upgrade as an app's requirements grow
  - **SQL**
    - **MySQL –** Kubernetes & Cloud Foundry
    - **Postgres –** Kubernetes & Standalone
    - **SQL Server** - Standalone

```
SELECT MAX(Price) AS LargestPrice
FROM Products;
```



VMware SQL with Postgres Offering

Why Postgres and Kubernetes

Simplified provisioning and maintenance

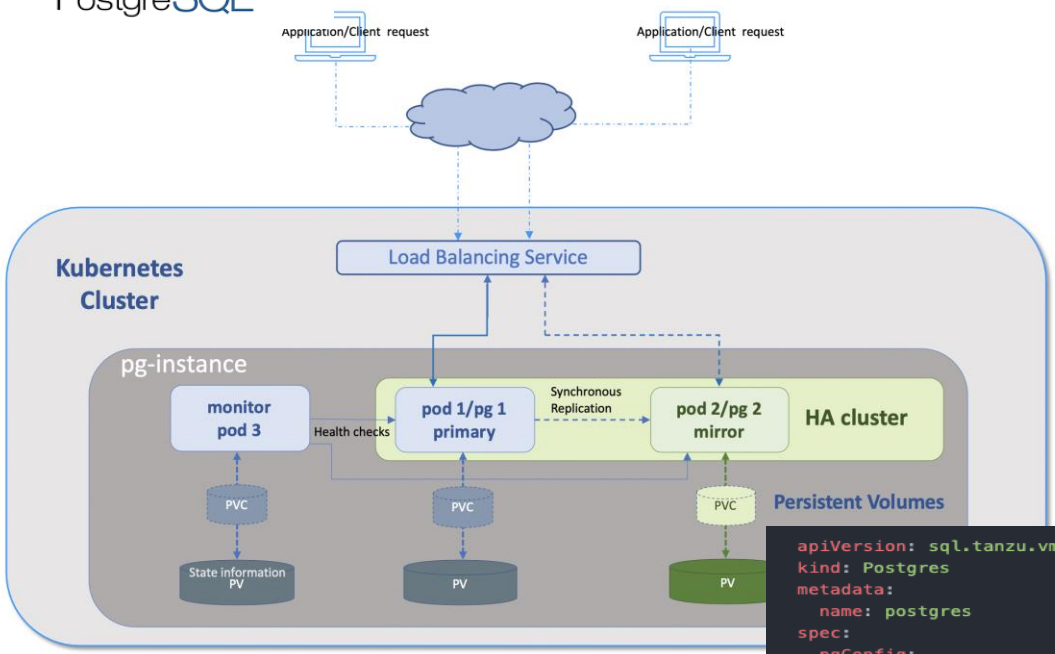
- Self-provisioning on-premises and in the public cloud.
- Includes Certified software packages, updates, bug fixes, and security patches
- 24x7 technical support

Availability and Scalability

- Configure Postgres deployment with HA with automated failover
- Take backups & restore the Postgres database via pgBackRest
- Scale & update deployed instances
- Pg_auto_failover for robust replication

VMware SQL for Kubernetes

- Day 2 operational Coverage
- HA, Backup & Restore, Certification Rotation
- Secure by defaults
- Designed to manage a fleet of instances within a Kubernetes environment
- In combination with Data Management for VMware Tanzu, manage a fleet of DBs anywhere




```
apiVersion: sql.tanzu.vmware.com/v1
kind: Postgres
metadata:
  name: postgres
spec:
  pgConfig:
    dbname: postgres-sample
    username: pgadmin
    appUser: pgappuser
  postgresVersion:
    name: postgres-14
  serviceType: ClusterIP
  seccompProfile:
     type: RuntimeDefault
     highAvailability:
       enabled: true
     logLevel: WARN
     backupLocation:
       name: backuplocation-sample
  storageClassName: standard
  storageSize: 800M
  cpu: "0.8"
  memory: 800Mi
  dataPodConfig:
   resources:
     requests:
       cpu: 800m
       memory: 800Mi
```

VMware SQL with MySQL for Kubernetes

Why MySQL and Kubernetes

MySQL engine

- 2nd most popular DB engine in the world

Simplified provisioning and maintenance

- Self-provisioning on-premises and in the public cloud.
- Includes Certified software packages, updates, bug fixes, and security patches
- 24x7 technical support

VMware SQL with MySQL for Kubernetes

- Day 2 operational Coverage
- HA, Backup & Restore, Certification Rotation
- Major improvements to group Replication
- Secure defaults
- Designed to manage a fleet of instances within a Kubernetes environment
- In combination with Data Management for VMware Tanzu, manage a fleet of MySQL anywhere

```yaml
apiVersion: with.sql.tanzu.vmware.com/v1
kind: MySQL
metadata:
  name: mysql-sample
spec:
 mysqlVersion:
   name: mysql-8.0.26
 storageSize: 1Gi
 storageClassName: standard
 serviceType: LoadBalancer
 highAvailability:
   enabled: true
 resources:
   mysql:
     requests:
       cpu: 750m
       memory: 500Mi
```

Greenplum

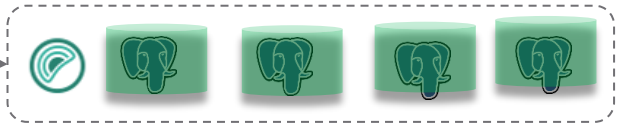
Big Data

Analytics

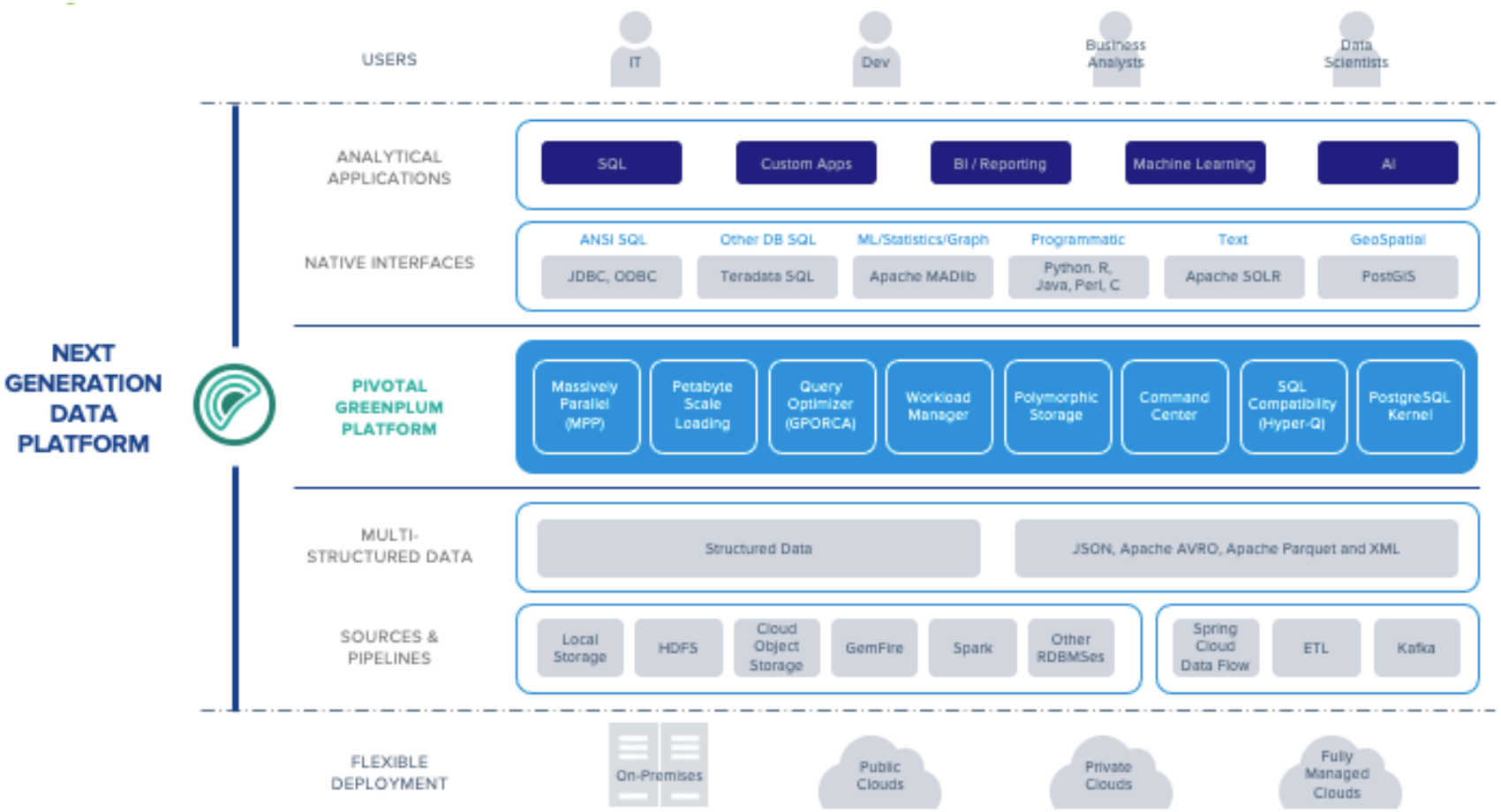
Greenplum

Use Cases

```
SELECT * FROM gptext.index
(
  TABLE(SELECT *
        FROM messages
        SCATTER BY distrib_id),
        'mytest.articles'
);
```



- Scalable Query access
  - **SQL** compliant
  - Massively Parallel Processors (MPP)
  - Postgres clustered instances
  - Petabyte scale
- Flexible **data science** framework for data learning and access
  - Embedded **Machine Learning**
  - Geospatial
  - Graph and Text **Analytics**
  - pl/container for functions (ex: Python & R functions)
  - **Apache Madlib** – embedded open-source SQL based mathematical, statistical, graph and machine learning library
- **Aggregate** access/integrate data from external data sources
  - Connectivity with Message Brokers, Spark, S3, Hadoop and more.



Find anyone who works at 'VMware' and know each other 'directly' and whose names sound like 'Peter' or 'Pavan' and have withdrawn an amount > $200 within 24 hours at an ATM less than 2 KM from reference latitude and longitude.

```
CREATE FUNCTION get_people(person1 text, person2 text, amount int, duration int, longit float,latitude float) RETURNS int
AS $$
declare
linkchk integer; v1 record; v2 record;
begin
 execute 'truncate table results;';
 for v1 in select distinct a.id,a.firstname,a.lastname,amount,tran_date,c.lat,c.lng,address,a.description,d.score from people a,transactions b,location c,
    (SELECT w.id, q.score FROM people w, gptext.search('gpadmin.public.people' , 'VMware') q
     WHERE (q.id::integer) = w.id order by 2 desc) d
     where soundex(firstname)=soundex($1) and a.id=b.id and amount > $200 and (extract(epoch from tran_date) - extract(epoch from now()))/3600 < $4

and st_distance_sphere(st_makepoint($5, $6),st_makepoint(c.lng, c.lat))/1000.0 <= 2.0 and b.locid=c.locid and a.id=d.id
 loop
   for v2 in select distinct a.id,a.firstname,a.lastname,amount,tran_date,c.lat,c.lng,address,a.description,d.score from people a,transactions b,location c,
    (SELECT w.id, q.score FROM people w, gptext.search(TABLE(SELECT 1 SCATTER BY 1), 'gpadmin.public.people' , 'Pivotal', null) q
     WHERE (q.id::integer) = w.id order by 2 desc) d
     where soundex(firstname)=soundex($2) and a.id=b.id and amount > $3 and (extract(epoch from tran_date) - extract(epoch from now()))/3600 < $4
    and st_distance_sphere(st_makepoint($5, $6),st_makepoint(c.lng, c.lat))/1000.0 <= 2.0 and b.locid=c.locid and a.id=d.id
 loop
     execute 'DROP TABLE IF EXISTS out, out_summary;';
     execute  'SELECT madlib.graph_bfs(''people'',''id'',''links'', v1.id,''out'');';
     select 1 into linkchk from out where dist=1 and id=v2.id;
     if linkchk is not null  then
          insert into results values (v1.id,v1.firstname,v1.lastname,v1.amount,v1.tran_date,v1.lat,v1.lng,v1.address,v1.description,v1.score);
          insert into results values (v2.id,v2.firstname,v2.lastname,v2.amount,v2.tran_date,v2.lat,v2.lng,v2.address,v2.description,v2.score);
      end if;
    end loop;
 end loop;
 return 0;
end
$$ LANGUAGE plpgsql;
-- Call the function now
select get_people('Pavan','Peter',200,24,103.912680, 1.309432) ;
```

Greenplum Fuzzy String Match function **Soundex()** to know if people name sounds like 'Pavan' or 'Peter'

GPText.search() function is used to know if both people work at 'Vmware'

Amount > $200

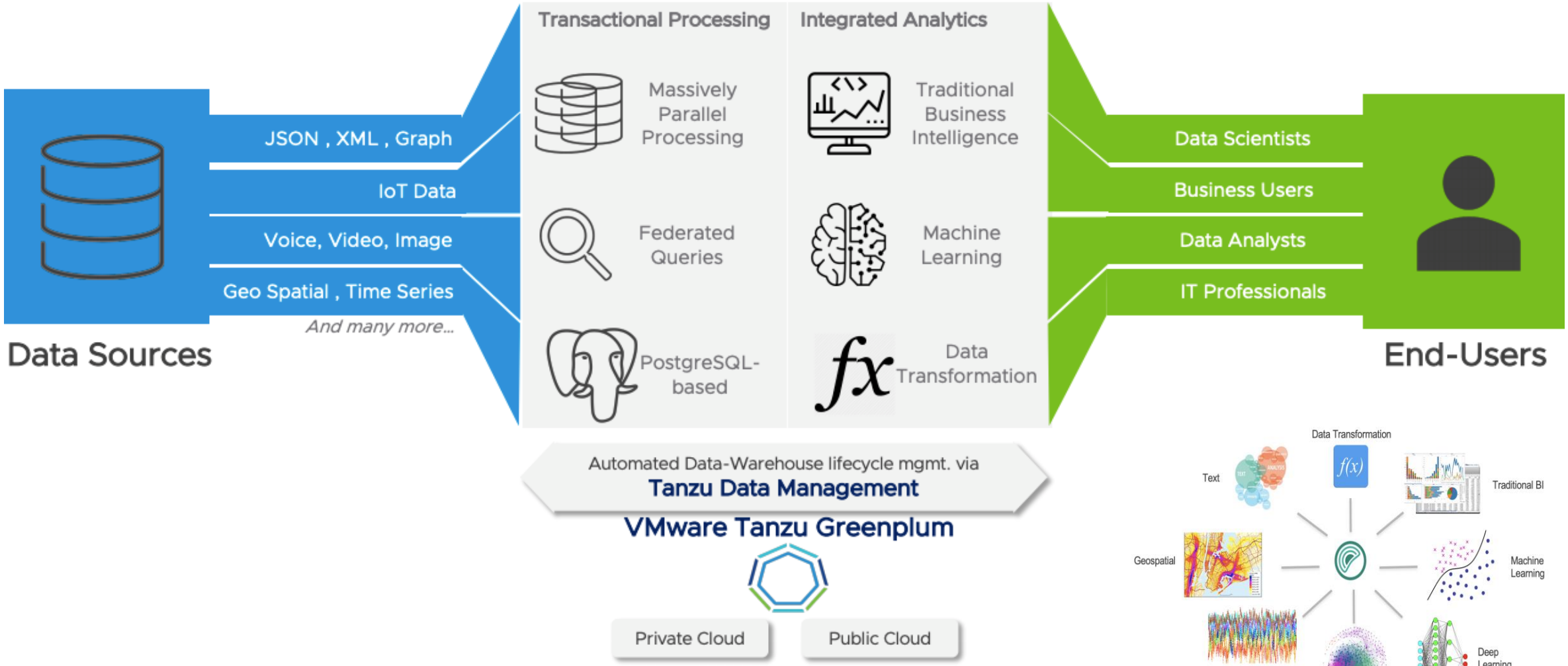
Greenplum and **Apache MADlib** BFS search to know if there are direct or indirect links between people

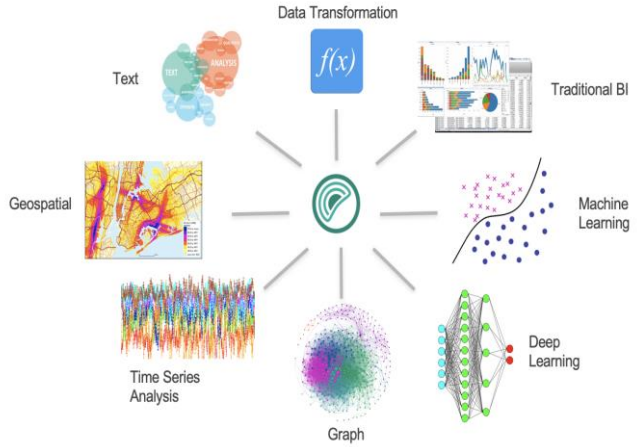
Greenplum **Time functions** to calculate difference in amount withdrawn time < 24 hours

Greenplum POSTGIS functions **st_distance_sphere() and st_makepoint()** calculate distance between ATM location and reference latitude, longitude < 2 KM

Greenplum: Federated database

Traditional and Advanced Analytics In-Database





Greenplum versus VMware Greenplum

Open source versus commercial features

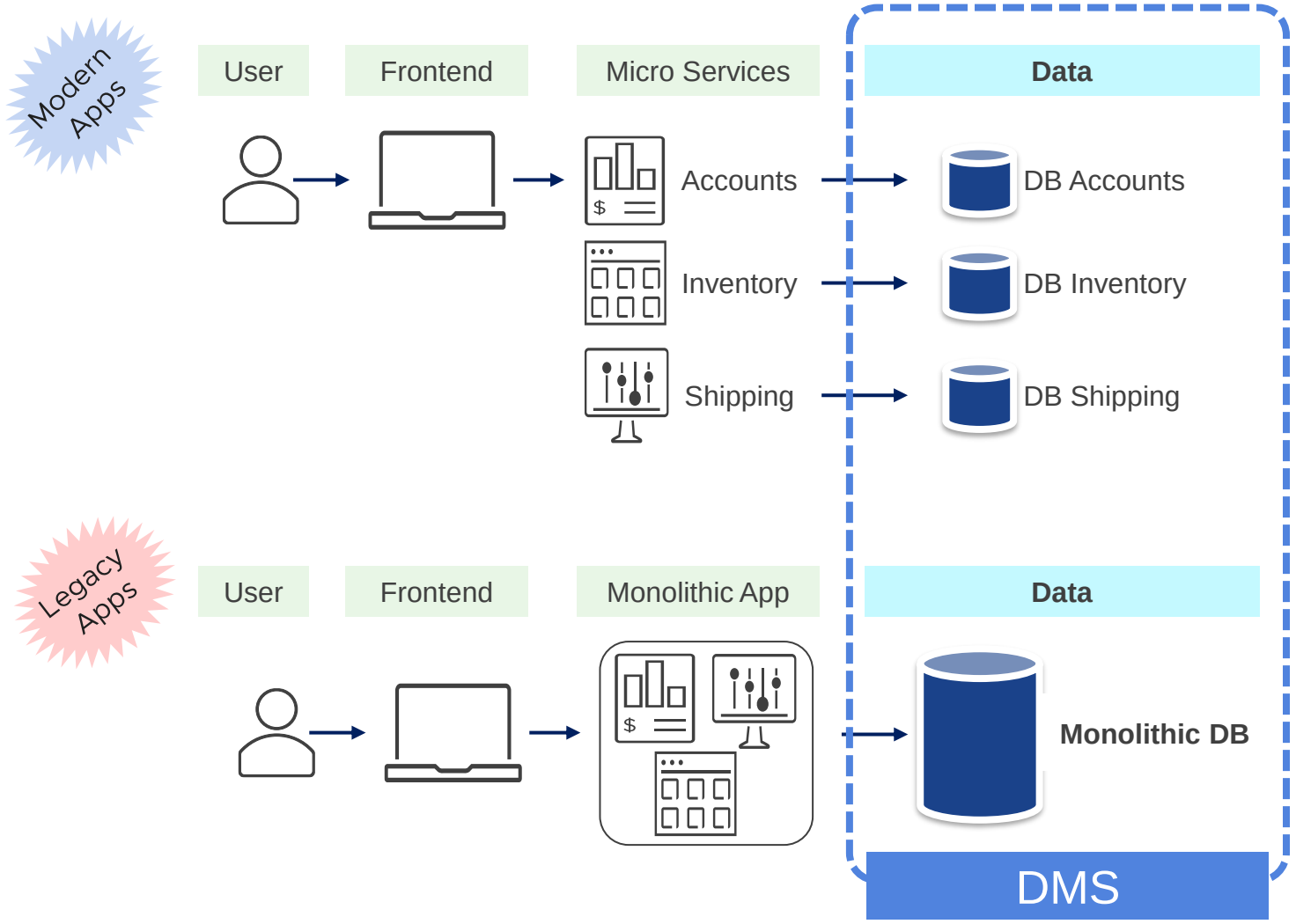
	Open Source GPDB	VMware Greenplum
Greenplum Database	X	X
Platform Extension Framework (PXF)	X	X
PL/Container	X	X
PL/Java	X	X
PL/Perl	X	X
PL/pgsql	X	X
PL/Python	X	X
PL/R	X	X
MADlib	X	X
PostGIS		X
24x7x365 Follow-the-Sun Support		X
GPText (Solr)		X
Greenplum Stream Server (GPSS / gpkafka)		X
Greenplum Data Copy (gpcopy)		X
Greenplum Backup Restore Manager		X
Data Domain DeDupe		X
Progress DataDirect		X
Spark Connector		X
Data Science Packages		X
Greenplum Command Center		X
Certified Partners		X

Tanzu Data Management Services

Data has history: from Legacy to Microservices

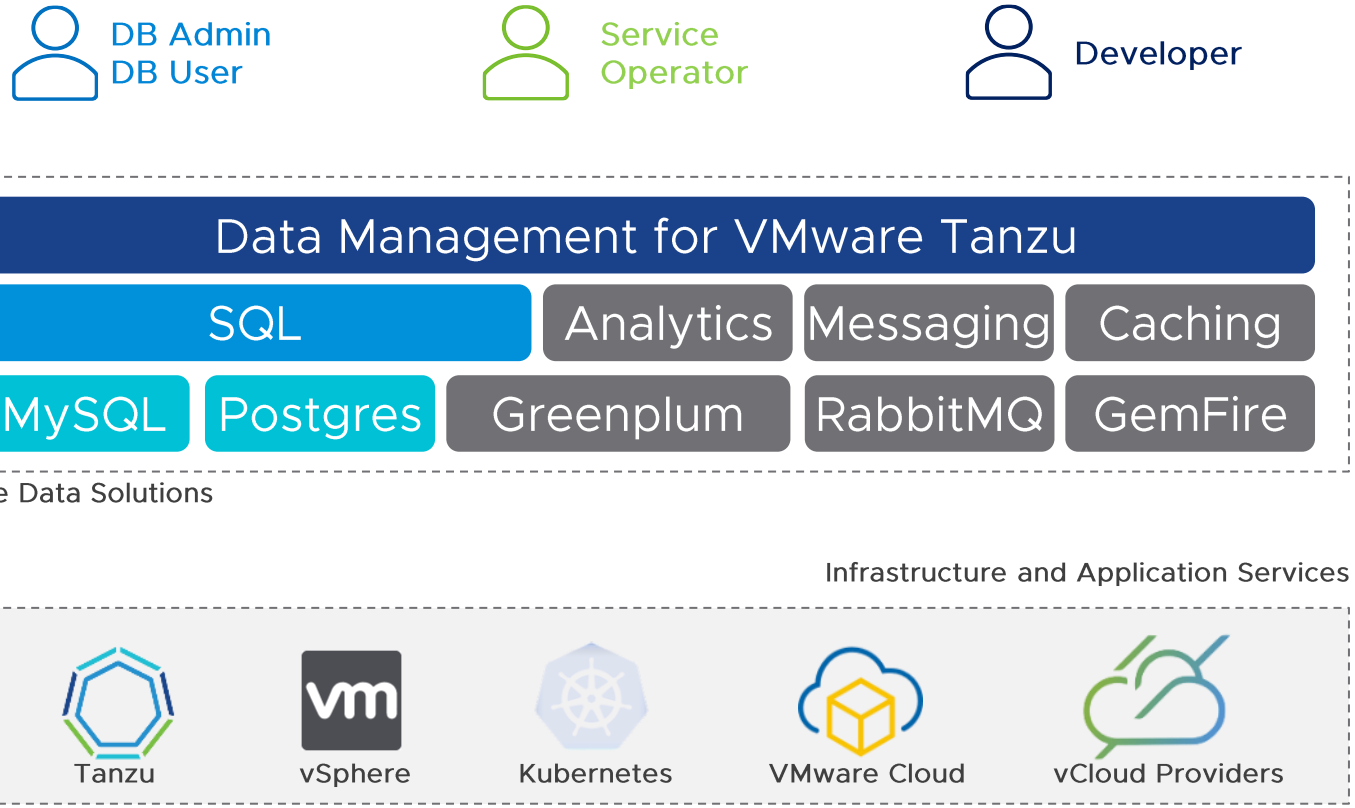
Co-existence of monolithic & modern databases leads to rapid data growth

- Modern apps are independently deployable and scalable
- Proliferation of many different types of databases
- Leads to co-existence of both modern and legacy data technologies
- Self-management of data and databases poses significant operational overhead, risks the data layer inhibiting developer productivity



Tanzu Data Management Services

A Data Experience That Matches the Public Cloud



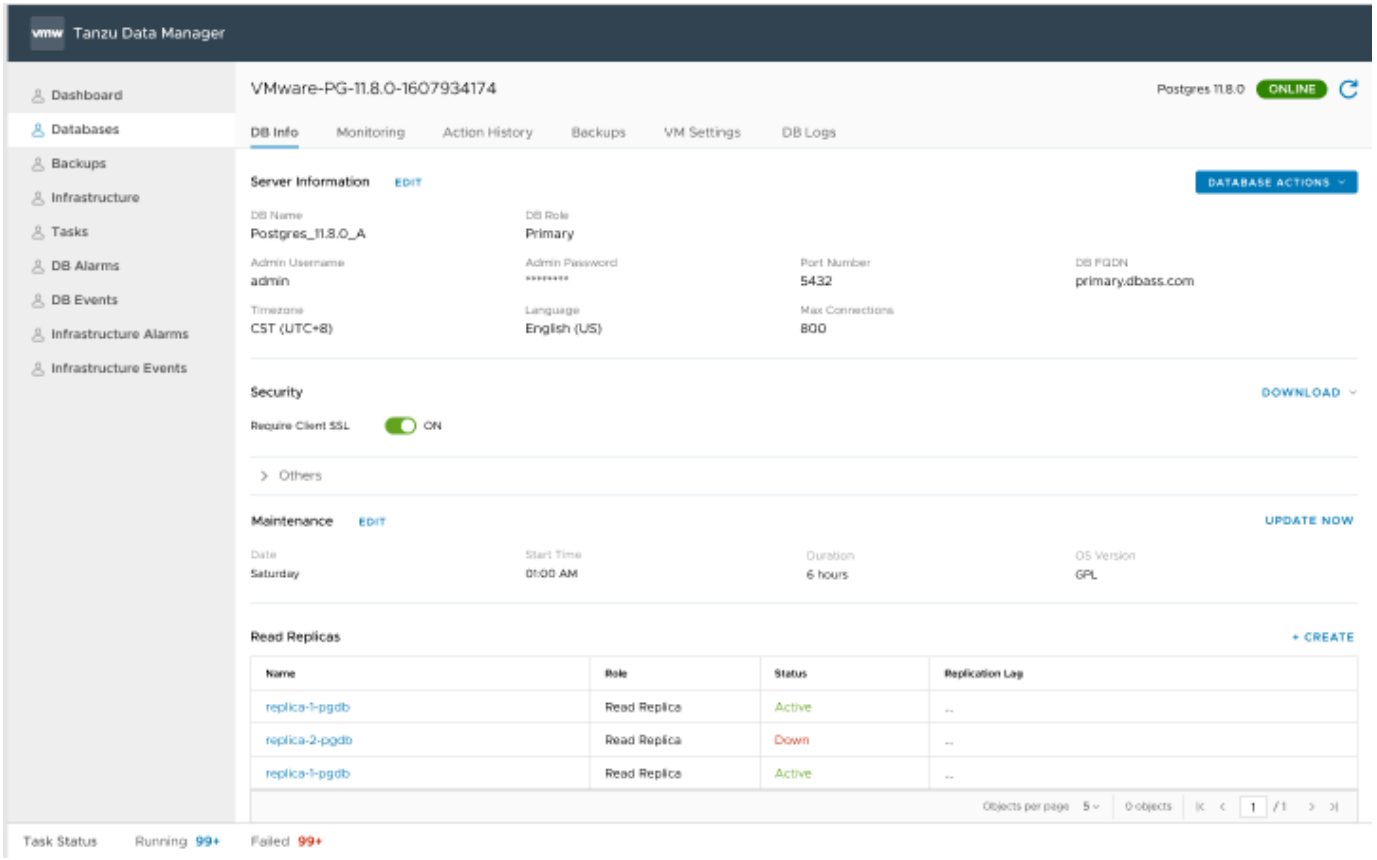
Value Proposition

- **Deliver** a set of Foundational Data Services with cloud- native abstractions

VMWare Tanzu	Amazon AWS
Tanzu SQL	RDS
Tanzu GemFire	ElastiCache
Tanzu Greenplum	RedShift
Tanzu RabbitMQ	SQS / AmazonMQ

- **Operator Efficiency** - an integrated control plane that offers foundational data services
- **Portability**: Any platform: vSphere, k8s, VCPP, VMC
- **Security & Availability:** Fleet management, HA, day 2 ops, backups
- **Drive cost optimization** by moving maintenance to best/low cost

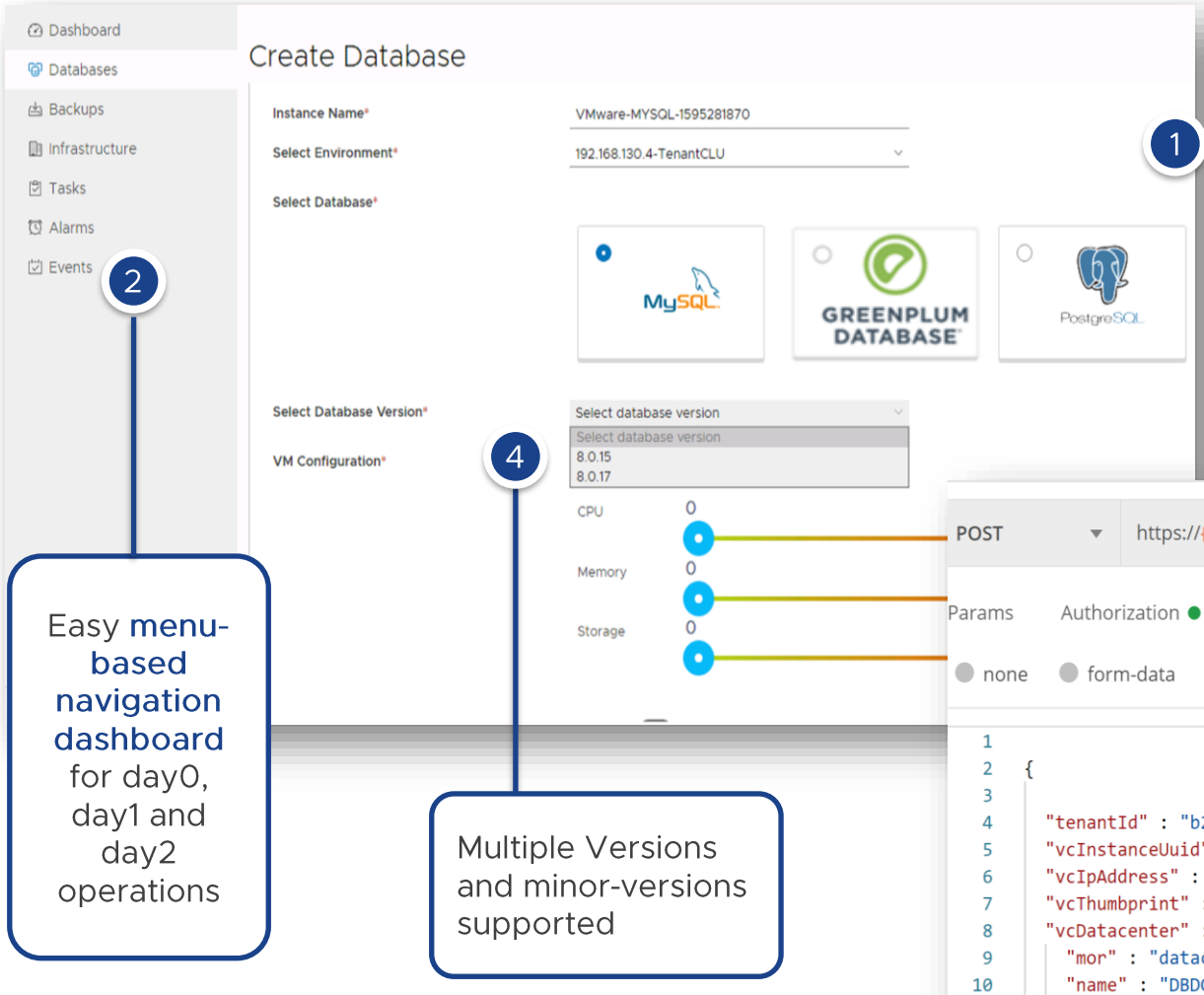
Single Control Plane To Simplify Management Of Tanzu Data Products



- A Fleet Management toolkit that allows customers or providers to host their own Data-as-a-Service on their own infrastructure or cloud accounts
- Self-service with on demand day-0, day-1, and day-2 automations, including security patching and minor updates
- Self-service and multi-tenancy provided via UI and via REST API.
- Very similar to Amazon RDS (today), (coming) AWS SQS and AWS ElastiCache

Ease of Use

Simple and Intuitive UI, UX and API to tackle advanced tasks with ease

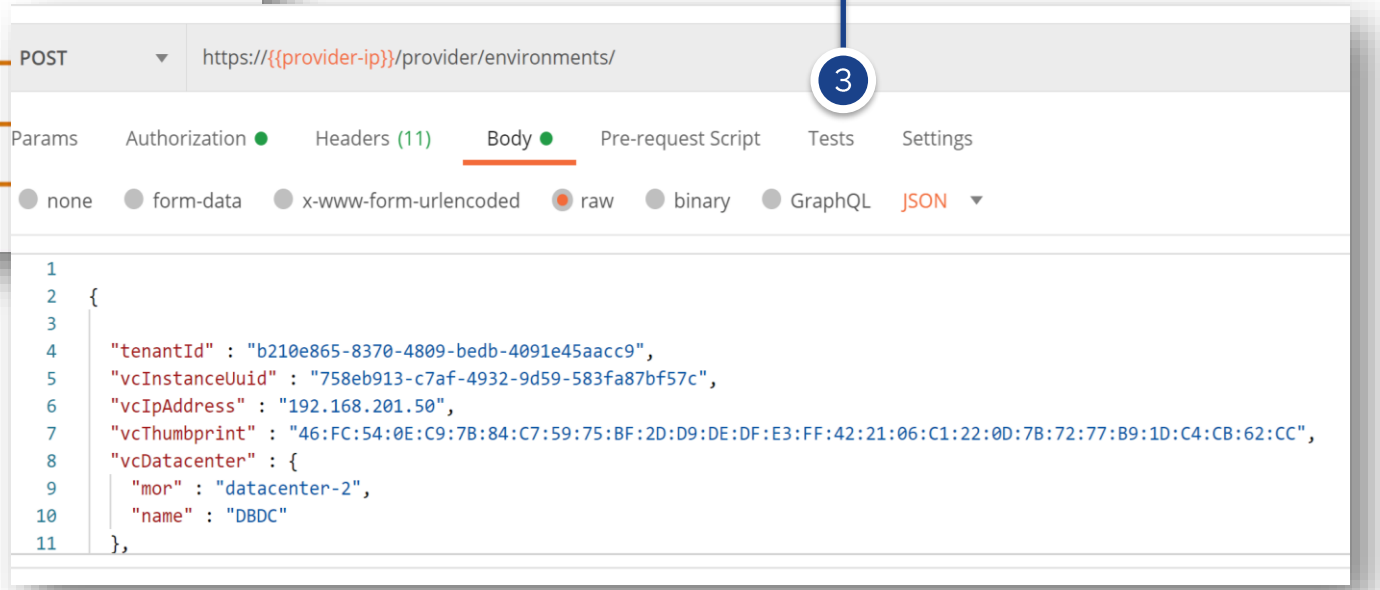


1

Choice to use **One-click database operations** using preconfigured default settings, ranging from Provisioning to Cloud Archive, including read replicas

- **All major database operations available via DB Admin and DB Service Admin APIs**
- Can be integrated with any other DevOps framework

3



2

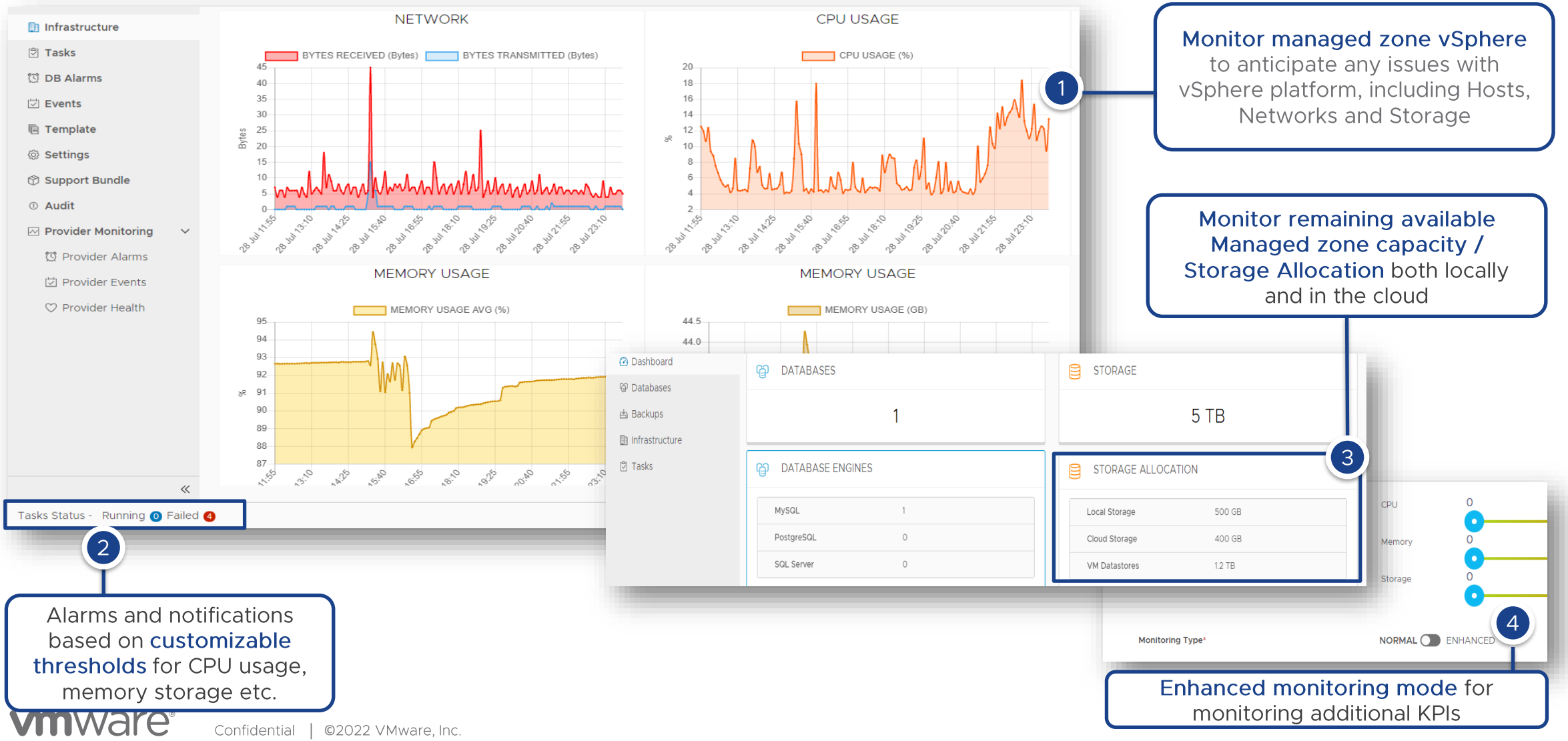
Easy **menu-based navigation dashboard** for day0, day1 and day2 operations

4

Multiple Versions and minor-versions supported

Visibility

Monitoring capabilities for database and managed zone for performance insights



Monitor managed zone vSphere to anticipate any issues with vSphere platform, including Hosts, Networks and Storage

Monitor remaining available Managed zone capacity / Storage Allocation both locally and in the cloud

Alarms and notifications based on **customizable thresholds** for CPU usage, memory storage etc.

Enhanced monitoring mode for monitoring additional KPIs

Feature Highlights of Tanzu Data Management

For Users, via Self-Service Portal or API

Provisioning	• Provision different flavors and configurations of SQL Engines • Multitenant (Organizations and Users)
Patching	• Plan patch & updates (including OS), on demand or as scheduled
Backup	• Backup on demand or as scheduled • Archival to Cloud policy
Secure	• Encryption-at-rest • 2way TLS in between end-points, cert management
Replication	• Replicate (Cold/Hot Replicas or Read Replicas) on same / remote Managed Zone
Monitoring	• Monitor SQL Engine, Operating System, Networking, I/O etc. • DB Profiling • Observability and Log Consolidation

Technologies

MySQL

PostgreSQL

GREENPLUM DATABASE

GemFire

RabbitMQ

Thank You