Meet the demands of big data analytics with the in-memory speed of Aerospike

Real-time analytics on IBM SoftLayer

Skylab Vanga
Chelliah Pethuru Raj

October 07, 2015

Today, people want cloud environments to host high performance, enterprise class, and real-time workloads to meet varied expectations of individual users and businesses. Aerospike is an open source, real-time NoSQL database and key-value store that provides in-memory performance for big data and context-driven applications that must sense and respond "right now" at a fraction of previous costs. Aerospike operates on a global scale with enterprise-grade reliability. Clouds are destined to be both commercially viable and capable of delivering process and data-intensive applications. This article showcases how the Aerospike platform on IBM SoftLayer® cloud will be a game changer for big data analytics and other demanding applications.

Increasingly, process and data-intensive applications are being used in cloud environments. The cloud concept is undergoing changes to be positioned as the one-stop solution for all kinds of IT requirements. With the current exponential growth in data collection, high-quality data analysis algorithms are abounding. Cloud-induced infrastructure optimization is increasing through many mechanisms such as:

- Consolidation
- Centralization
- Federation
- Virtualization
- Containerization
- Commoditization
- Industrialization

We have policy-based and software-defined abstraction, automation, standardization, and simplification galore; these provide adaptive, instant-on, and optimized infrastructures (servers, storage, and networking elements). To host and run in elastic clouds elegantly, and to fulfill the varied needs of the big data world, we have prominent IT solutions: versatile in-memory computing, NoSQL and NewSQL databases, and parallel file systems.
Aerospike is an open source database built from the ground up to push the limits of flash memory, processors, and networks. It was designed to operate with predictable low latency at high throughput with uncompromising reliability — both high availability and ACID guarantees. Aerospike is:

- An in-memory, NoSQL database and key-value store that runs at wire speed.
- Able to substantially simplify your workloads, as there is no need to incorporate the logic for sharding and cluster changes.
- A game-changing database solution that takes away the worry about data loss or downtime.
- Operating at in-memory speed and global scale with enterprise-grade reliability.
- Ideal for real-time big data or context driven applications that must sense and respond right now.

In this article, learn how to migrate Aerospike to the IBM SoftLayer cloud environment and configure it to meet your needs. A sample application showcases Aerospike’s capabilities in harnessing data-intensive workloads.

The need for in-memory computing for real-time analytics

The Systems, Applications and Products in Data Processing (SAP) SE corporation has championed its HANA platform to run an entire company, encompassing:

- Mission-critical transactional applications such as ERP and CRM.
- Analytic needs previously handled by separate database management systems (DBMSs) underpinning data warehouses and data marts.

IBM has BLU Acceleration® for DB2®, which combines both columnar compression and in-memory processing to accelerate analytics. Teradata introduced an Intelligent Memory feature that automatically moved data that's most often queried into RAM to ensure the fastest-possible query response.

Aerospike undoubtedly leads the pack of products that facilitate in-memory NoSQL databases. There are significant technical differences between SAP’s all-in-memory HANA platform and Aerospike, which touts its in-memory options deployed non-disruptively (without ripping and replacing existing applications). Our focus is to explore the need for speed in the context of business. Several businesses claim that the need for in-memory speed is real and the subsequent business benefits can be enormous.

About Aerospike

Identical Aerospike servers scale out to form a shared-nothing cluster that transparently partitions data and parallelizes processing across nodes. Nodes in the cluster are identical; you can start with two and just add more hardware. The cluster scales linearly.

Use cases

Aerospike is a key-value store that is popularly used as a cache, or with persistence, as a context store. The context store could be used for:
• A server-side session store
• A cookie store
• A device ID store
• ID mapping
• User preferences or user profiles to make real-time recommendations and personalize the user experience on web portals
• E-commerce
• Travel sites

Aerospike is also the database of choice for real-time bidding platforms across display advertising, mobile, video, social media, gaming, native ads, and Internet TV.

For big data applications
As people visit websites and use mobile applications to click, swipe, or like, they leave a trail of big data that includes: page views and sensor data from things they wear, such as activity trackers; and things they own, such as light, smoke, and temperature sensors in their homes. New generation applications are using this data or context to anticipate needs and make predictions.

For context-driven applications
Context driven applications include digital advertising, multichannel marketing, email marketing, dynamic content serving, Identity Management, cross channel support, and loyalty management platforms. Commonly known applications include: travel portals, e-commerce portals with dynamic product pricing and inventory management, e-commerce search and personalized product recommendations, cyber security, fraud detection, and similar applications.

Aerospike for big data analytics
Today's web scale enterprise-grade application architecture typically consists of Aerospike behind a web application tier and in front of a legacy DBMS or HDFS cluster. Large volumes (i.e., petabytes) of archive and historical data are stored on low-cost rotational drives. Insights or segments from “HDFS analytics” are periodically moved into Aerospike. These insights are then combined with terabytes of real-time data stored in RAM or flash storage on Aerospike. Applications use this rich user context as well as “hot analytics” (i.e., massively distributed aggregations) in Aerospike to make the best decisions and recommendations.

You can also configure Aerospike to store the latest data and automatically age out or expire older data. Figure 1 shows an architecture for Aerospike-based big data analytics.

Figure 1. Architecture for Aerospike-based big data analytics
Aerospike is a row store in which data is stored in records (key-value) and grouped into sets and namespaces (tables). Each 128K-2MB record can contain values (map, list, integer, string, blob types) that can be changed immediately.

**Sample application**

Our sample application shows that Aerospike data structures on top of a key-value store are an effective way to write applications with Aerospike as the only database. The sample describes the design and implementation of a Twitter-like application. The code is easy to follow but substantial enough to use as a foundation to leverage Aerospike’s technology. You can also use the sample as a seed application for expansion.

**Minimum requirements for Aerospike in IBM SoftLayer cloud**

**Memory (RAM)**
You will need 4GB of RAM. Because indexes are stored in memory itself, the amount of memory will limit the number of rows the hardware can store. Aerospike is very memory efficient and each row (object or record) requires only 64 bytes of memory for the index. Each GB of memory can index 16 million rows, and a 4GB memory configuration can index only 64 million objects. For development purposes, you can provision as little as 2GB of RAM.

**CPU**
You will need one quad-core CPU. Though there is no direct dependency on the CPU, you might find that CPUs are saturated quickly with system interrupts. Our example uses 8GB of RAM because we are going to install an Aerospike instance and Aerospike Management Console on a single node. (There is an option to install Aerospike and the Management Console in different nodes, so we could choose two nodes if desired.) You need a VM from the IBM SoftLayer cloud through the SoftLayer portal. The minimum configuration of VM is 8GB of RAM, 25GB disk storage, and two cores. This works across all the major Linux distributions. Our proof of concept (PoC) sample application is implemented using CentOS 64 bit.

**Host name**
aerospikepoc.softlayer.com
Address: 10.76.60.39 / 184.173.49.2

**User**
root / xxxxxx

**Installing Aerospike**

To install Aerospike, open your VM session then follow the steps below.

1. From the command line, enter the commands in Listing 1.
Meet the demands of big data analytics with the in-memory speed of Aerospike

### Listing 1. Get aerospike.tgz file

```bash
cd /usr
wget -O aerospike.tgz 'http://aerospike.com/download/server/latest/artifact/el6'
```

Your screen should show HTTP request and connection messages, as in the example below.

![HTTP request and connection messages](image)

2. You should see the aerospike.tgz file in the host in the /usr directory, as shown in Figure 2.

**Figure 2. aerospike.tgz file in the /usr directory**

![aerospike.tgz file in the /usr directory](image)

**Before** you install Aerospike, you must:

1. Turn Off SELinux. From the command line, enter `SELINUX=disabled`, as shown below.

   ```bash
touch /etc/selinux/config
sudo sed -i 's/SELINUX=enabled/SELINUX=disabled/g' /etc/selinux/config
```

2. Turn off IPTables. From the command line, enter `chkconfig iptables off`, as shown below, then enter `iptables stop`.

3. Turn on NTP. From the command line, enter the commands shown in Listing 2.

**Listing 2. Turn on NTP**

```bash
sudo /sbin/chkconfig ntpd on
sudo ntpdate pool.ntp.org
/etc/init.d/ntpd start
```
4. Go to the Aerospike directory. From the command line, enter the `cd /usr` command.

```
root@aerospike@ /usr/aerospike-server-community-3.4.1-el6 ](1)
```

5. Extract the aerospike.tgz. From the command line, enter the `tar -xvf aerospike.tgz` command. You will see a brief list of the tools, files, and license in the package.

```
root@aerospike@ /usr/aerospike@ /usr/aerospike-server-community-3.4.1-el6 ](1)
```

Install and start Aerospike server and tools

To install the Aerospike server and the tools packages:

1. From the command line, enter the commands in Listing 3.

**Listing 3. Start Aerospike installation**

```
cd /usr/aerospike-server-community-3.4.1-el6
./asinstall
```

You will see messages on the screen as Python, the license, files, and groups are installed.

```
root@aerospike@ /usr/aerospike-server-community-3.4.1-el6 ](1)
```

After the installation is complete, the server should be ready for use with the default configuration file. (See the directory structure for more details.)

2. Aerospike uses several directories to store tools, system files, and data files. To get a description of the directories in use, or to manually manage files (they are typically managed through Aerospike tools), enter `/opt/aerospike` from the command line.

The aerospike directory is created and managed by the Aerospike packages when installed through Linux package management. It contains several subdirectories, some of which were created by the tools package and some of which were created by the server package and maintained during runtime of the server. As shown below, bin, data, doc, examples, lib, cmd, sys, and usr are included.
To start the Aerospike server, from the command line enter the commands in Listing 4.

**Listing 4. Start Aerospike server**

```
cd /etc/init.d
./aerospike start
```

On the screen, you will see a list similar to the following: abrt-ccpp, abrtd, abrt-oops, acpid, aerospike, atd, auditd, blk-availability, cpuspeed, crond, functions, haldaemon, halt, htcacheclean, HTTPd, ip6tables, iptables, and irqbalance.

To check the server status, enter `/aerospike status` from the command line.

Install the Aerospike client Java package

To install the Aerospike client Java package, you can use the WinSCP tool to copy the aerospike-client-java-3.0.33.tgz package to the `/usr` directory. To extract the package, enter `tar -xvf aerospike-client-java-3.0.33.tgz` from the command line.

Install the Aerospike Management Console manual

Download the file containing the Aerospike Management Console (AMC) manual from the Aerospike documentation website, then use the WinSCP tool to move it to the `/usr` directory by entering the commands in Listing 5.

**Listing 5. Put documentation in `/usr` directory**

```
cd /usr
yum install aerospike-amc-community-3.5.0-e15.x86_64.rpm
```

You will see the plugins that are being installed and dependencies that are resolved.
Start the AMC server

To start the AMC server, go to the amc directory and from the command line, enter the commands in Listing 6.

**Listing 6. Start AMC server**

```
cd /etc/init.d
./amc start
```

To check the status of the AMC server, enter `./amc status` from the command line. The output shows AMC is running.

Now you can start the Aerospike server by entering the commands in Listing 7.

**Listing 7. Start Aerospike server**

```
cd /etc/init.d
./aerospike start
./aerospike status
```

Thereafter, use the URL `http://184.173.49.2:8081` in your browser; this will open the window shown in Figure 3.

**Figure 3. Connecting to a node**
Enter your Host Name or IP then click **Connect** to see the Aerospike dashboard, which is shown in Figure 4 and Figure 5.

**Figure 4. Aerospike dashboard**

![Aerospike dashboard](image)

**Figure 5. Aerospike dashboard, continued**

![Aerospike dashboard, continued](image)

**Using Aerospike**

At the top of the Aerospike main dashboard screen, you can select from the following choices (Enterprise Edition will have additional pages):

- Dashboard
Click the **Change Cluster** link in the upper right of the window to monitor another cluster.

You can change the time interval that is being graphed by changing the **Snapshot for last** value. The **Statistics** board in Figure 6 shows the result.

**Figure 6. Statistics view**

![Statistics view](image)

Select **Definitions** at the top of the Aerospike Dashboard to see the definitions of a namespace, as shown in Figure 7.

**Figure 7. Definitions dashboard**

![Definitions dashboard](image)
You can use the System Monitor to check on the health of the clusters. From the VM command line, enter the `asmonitor` command. Sample output shows creation of an initial config file and one host in cluster: 10.76.60.39:3000.

![System Monitor Output](image)

The output shows the three types of monitors:

- **info**, which is the high-level information about the Aerospike cluster.

  ![Monitor Info](image)

- **latency**, which provides details about cluster throughput and latency.

  ![Monitor Latency](image)

- **printconfig**, which shows the configuration of the node (its actual output).

  ![Monitor Printconfig](image)

To leave the monitor section, enter the `exit` command from the VM command line.

**Adding test data into Aerospike**

To add test data to Aerospike, enter the commands in Listing 8.
Listing 8. Add test data to Aerospike

```bash
ascli put <ns> <set> <key> <record>
ascli put test testset testkey1 '{"name": "John"}'
```

Based on your data entered into the disk and the usage of RAM, the usage will be changed, as shown in Figure 8. The **Cluster Throughput (Hide)** will also show some variations.

Figure 8. Cluster disk usage

Administrative tasks

To start the Aerospike server, enter `sudo service aerospike start` from the command line.

To verify that Aerospike is running, enter `sudo service aerospike status` from the command line.

AMC server tasks

To start the AMC, enter `sudo /etc/init.d/amc start` from the command line.

To stop the AMC server, enter `sudo /etc/init.d/amc stop` from the command line.

To restart the AMC server, enter `sudo /etc/init.d/amc restart` from the command line.

To see whether or not the AMC server is up and running, enter `sudo /etc/init.d/amc status` from the command line.

To change the password, from the command line enter the command in Listing 9.

Listing 9. Change password

```bash
sudo /opt/amc/bin/reset_password [-u <username>] [-p <password>]
```
The options in this command are:

- `-u`: Optional username (only `admin` is allowed for now).
- `-p`: Optional new password. If not specified, the default password will be `admin`.

To see the server log, from the command line enter the commands in Listing 10.

**Listing 10. See the server log**

```bash
var/log/aerospike/aerospike.log
grep cake /var/log/aerospike/aerospike.log
```

**Conclusion**

Performing real-time analytics on voluminous, streaming, and multi-structured data is not easy; we need multifaceted solutions and platforms. Extracting actionable insights on big and fast data is a critical challenge for product vendors.

Aerospike is an innovative solution for simplifying and streamlining real-time analytics without compromising any technical requirements. With Aerospike, performance, availability, scalability, and security are fully ensured before you embark on the journey into real-time analytics on big data. The key differentiator with Aerospike is its NoSQL database in system memory, which allows real-time data analytics.
Related topic

- Learn about and download Aerospike.

© Copyright IBM Corporation 2015
Trademarks
(www.ibm.com/developerworks/ibm/trademarks/)