Getting Started with AWS

Analyzing Big Data

Getting Started with AWS: Analyzing Big Data

Copyright © 2014 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

The following are trademarks of Amazon Web Services, Inc.: Amazon, Amazon Web Services Design, AWS, Amazon CloudFront, Cloudfront, CloudTrail, Amazon DevPay, DynamoDB, ElastiCache, Amazon EC2, Amazon Elastic Compute Cloud, Amazon Glacier, Kinesis, Kindle, Kindle Fire, AWS Marketplace Design, Mechanical Turk, Amazon Redshift, Amazon Route 53, Amazon S3, Amazon VPC. In addition, Amazon.com graphics, logos, page headers, button icons, scripts, and service names are trademarks, or trade dress of Amazon in the U.S. and/or other countries. Amazon's trademarks and trade dress may not be used in connection with any product or service that is not Amazon's, in any manner that is likely to cause confusion among customers, or in any manner that disparages or discredits Amazon.

All other trademarks not owned by Amazon are the property of their respective owners, who may or may not be affiliated with, connected to, or sponsored by Amazon.

Table of Contents

Analyzing Big Data 1

Key AWS Services for Big Data 1

Setting Up 3

Sign Up for AWS 3

Create a Key Pair 3

Tutorial: Sentiment Analysis 5

Step 1: Create a Twitter Developer Account 6

Step 2: Create an Amazon S3 Bucket 6

Step 3: Collect and Store the Sentiment Data 7

Launch an Instance Using AWS CloudFormation 7

Collect Tweets Using the Instance 8

Store the Tweets in Amazon S3 9

Step 4: Customize the Mapper 10

Step 5: Create an Amazon EMR Cluster 11

Step 6: Examine the Sentiment Analysis Output 14

Step 7: Clean Up 15

Tutorial: Web Server Log Analysis 16

Step 1: Create an Amazon EMR Cluster 16

Step 2: Connect to the Master Node 18

Step 3: Start and Configure Hive 19

Step 4: Create the Hive Table and Load Data into HDFS 19

Step 5: Query Hive 20

Step 6: Clean Up 21

More Big Data Options 23

Related Resources 25

Analyzing Big Data with Amazon Web Services

The following tutorials show you ways to use Amazon Web Services to process big data:

- Tutorial: Sentiment Analysis (p. 5) — How to use Hadoop to evaluate Twitter data
- Tutorial: Web Server Log Analysis (p. 16) — How to query Apache web server logs using Hive

Key AWS Services for Big Data

With AWS, you pay only for the resources you use. Instead of maintaining a cluster of physical servers and storage devices that are standing by for possible use, you can create resources when you need them. AWS also supports popular tools like Hadoop, Hive, and Pig, and makes it easy to provision, configure, and monitor clusters for running those tools.

The following table shows how AWS can help address common big-data challenges.

Challenge	Solution
Data sets can be very large. Storage can become expensive, and data corruption and loss can have far-reaching implications.	Amazon S3 can store large amounts of data, and its capacity can grow to meet your needs. It is highly redundant and secure, protecting against data loss and unauthorized use. Amazon S3 also has an intentionally small feature set to keep its costs low.
Maintaining a cluster of physical servers to process data is expensive and time-consuming.	When you run an application on a virtual Amazon EC2 server, you pay for the server only while the application is running, and you can increase the number of servers — within minutes, not hours or days — to meet the processing needs of your application.