AWS Reference Architectures

- Amazon EC2
- Amazon EC2 Spot
- Amazon EMR
- Amazon S3
- Amazon CloudFront
- Amazon RDS

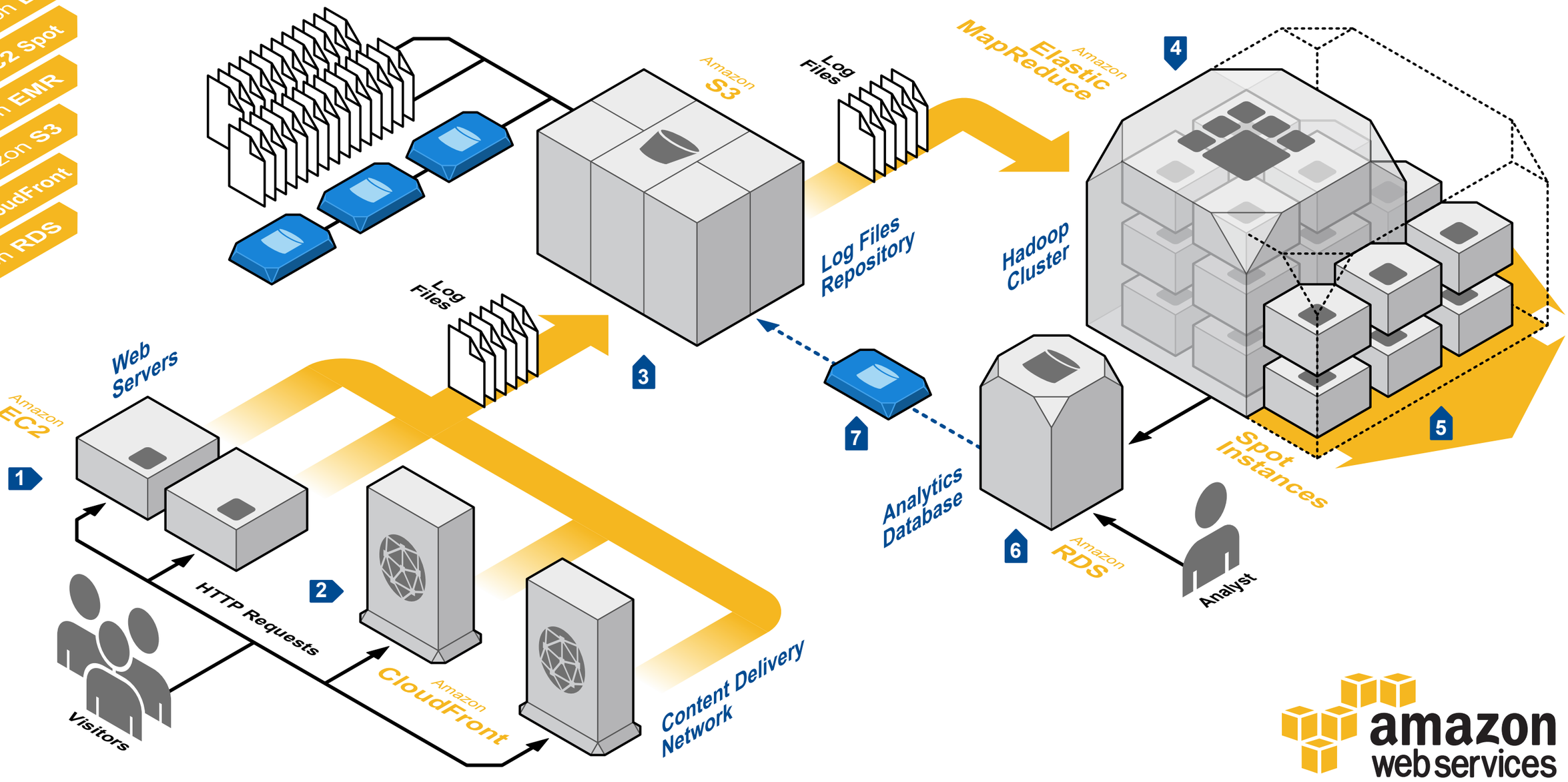
# WEB LOG ANALYSIS

Amazon Web Services provides services and infrastructure to build reliable, fault-tolerant, and highly available web applications in the cloud. In production environments, these applications can generate huge amounts of log information.

This data can be an important source of knowledge for any company that is operating web applications. Analyzing logs can reveal information such as traffic patterns, user behavior, marketing profiles, etc.

However, as the web application grows and the number of visitors increases, storing and analyzing web logs becomes increasingly challenging.

This diagram shows how to use Amazon Web Services to build a scalable and reliable large-scale log analytics platform. The core component of this architecture is Amazon Elastic MapReduce, a web service that enables analysts to process large amounts of data easily and cost-effectively using a Hadoop hosted framework.




amazon web services

## System Overview

**1** The web front-end servers are running on **Amazon Elastic Compute Cloud** (Amazon EC2) instances.

**2** **Amazon CloudFront** is a content delivery network that uses low latency and high data transfer speeds to distribute static files to customers. This service also generates valuable log information.

**3** Log files are periodically uploaded to **Amazon Simple Storage Service** (Amazon S3), a highly available and reliable data store. Data is sent in parallel from multiple web servers or edge locations.

**4** An **Amazon Elastic MapReduce** cluster processes the data set. **Amazon Elastic MapReduce** utilizes a hosted Hadoop framework, which processes the data in a parallel job flow.

**5** When **Amazon EC2** has unused capacity, it offers EC2 instances at a reduced cost, called the *Spot Price*. This price fluctuates based on availability and demand. If your workload is flexible in terms of time of completion or required capacity, you can dynamically extend the capacity of your cluster using **Spot Instances** and significantly reduce the cost of running your job flows.

**6** Data processing results are pushed back to a relational database using tools like *Apache Hive*. The database can be an **Amazon Relational Database Service** (Amazon RDS) instance. **Amazon RDS** makes it easy to set up, operate, and scale a relational database in the cloud.

**7** Like many services, **Amazon RDS** instances are priced on a pay-as-you-go model. After analysis, the database can be backed-up into **Amazon S3** as a database snapshot, and then terminated. The database can then be recreated from the snapshot whenever needed.