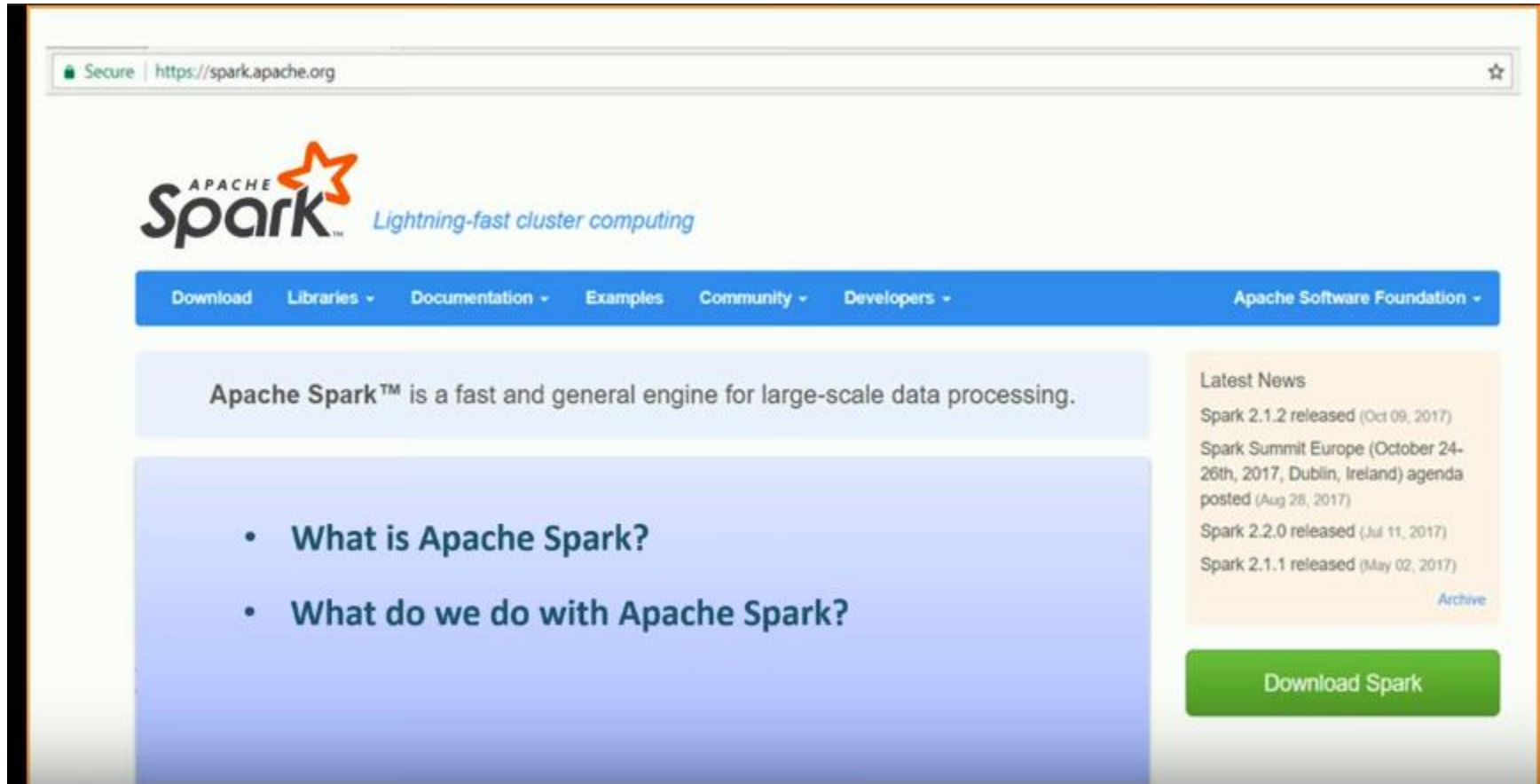


APACHE SPARK



Introduction



The screenshot shows the Apache Spark website at <https://spark.apache.org>. The page features the Apache Spark logo with the tagline "Lightning-fast cluster computing". A navigation bar includes links for Download, Libraries, Documentation, Examples, Community, Developers, and the Apache Software Foundation. The main content area states that Apache Spark is a fast and general engine for large-scale data processing. Below this, a blue box contains two bullet points: "What is Apache Spark?" and "What do we do with Apache Spark?". To the right, a "Latest News" section lists recent releases and events, including Spark 2.1.2, Spark Summit Europe, and Spark 2.2.0. A green "Download Spark" button is located at the bottom right of the page.

Secure | <https://spark.apache.org>

APACHE Spark™ Lightning-fast cluster computing

Download Libraries ▾ Documentation ▾ Examples Community ▾ Developers ▾ Apache Software Foundation ▾

Apache Spark™ is a fast and general engine for large-scale data processing.

- What is Apache Spark?
- What do we do with Apache Spark?

Latest News

Spark 2.1.2 released (Oct 09, 2017)

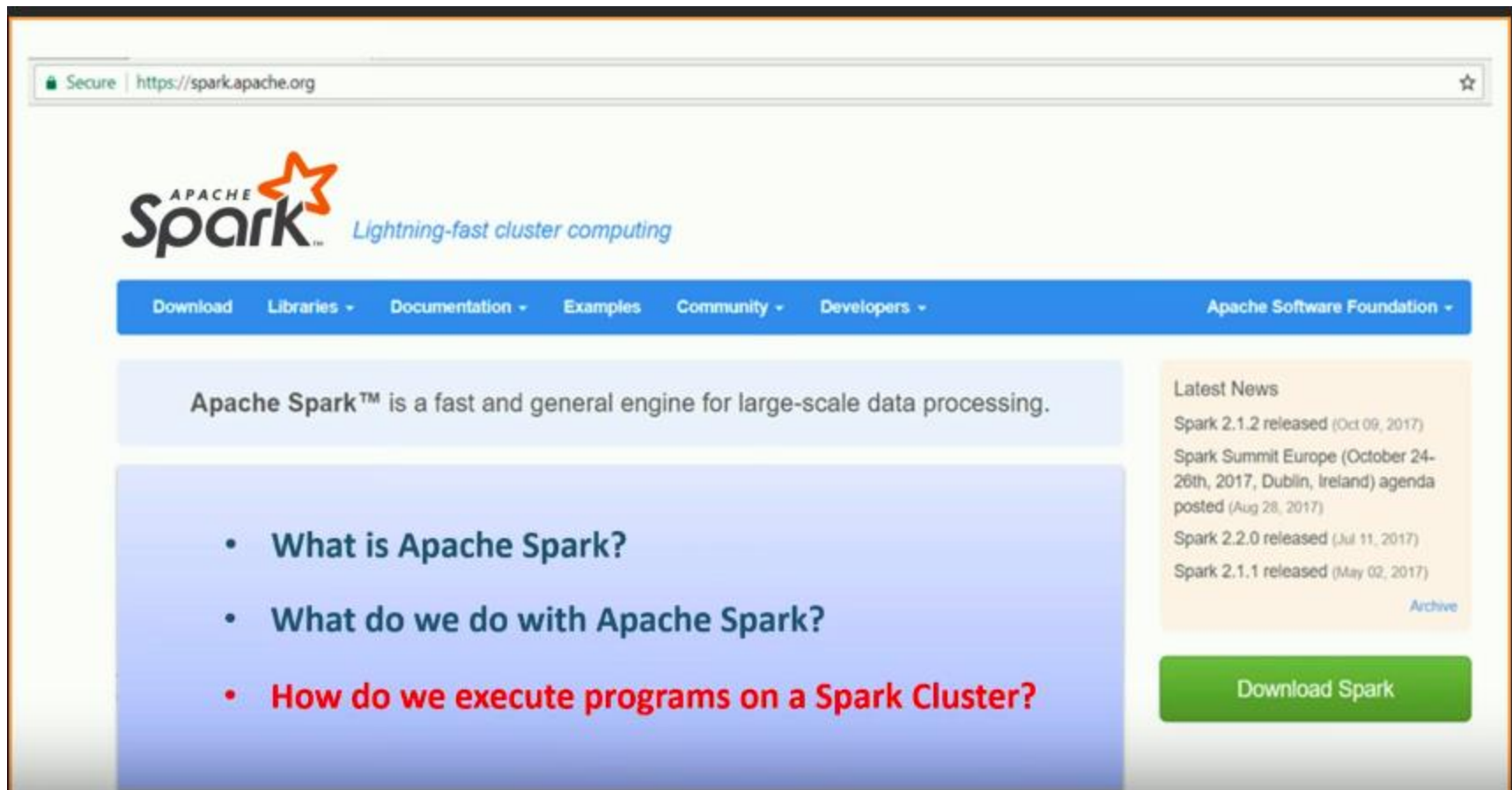
Spark Summit Europe (October 24-26th, 2017, Dublin, Ireland) agenda posted (Aug 28, 2017)

Spark 2.2.0 released (Jul 11, 2017)

Spark 2.1.1 released (May 02, 2017)

[Archive](#)

Download Spark



The screenshot shows the Apache Spark website. The browser address bar displays "Secure | https://spark.apache.org". The Apache Spark logo is prominently displayed, featuring the word "APACHE" in small letters above "Spark" in a large, bold font, with a stylized orange star above the "k". To the right of the logo, the tagline "Lightning-fast cluster computing" is written in a smaller, italicized font. Below the logo, a blue navigation bar contains links for "Download", "Libraries", "Documentation", "Examples", "Community", "Developers", and "Apache Software Foundation". The main content area has a light blue background. A text box states: "Apache Spark™ is a fast and general engine for large-scale data processing." Below this, a list of bullet points is shown: "What is Apache Spark?", "What do we do with Apache Spark?", and "How do we execute programs on a Spark Cluster?". To the right, a "Latest News" section lists recent updates: "Spark 2.1.2 released (Oct 09, 2017)", "Spark Summit Europe (October 24-26th, 2017, Dublin, Ireland) agenda posted (Aug 28, 2017)", "Spark 2.2.0 released (Jul 11, 2017)", and "Spark 2.1.1 released (May 02, 2017)". A green button labeled "Download Spark" is positioned at the bottom right of the page.

Secure | https://spark.apache.org

APACHE Spark™ Lightning-fast cluster computing

Download Libraries Documentation Examples Community Developers Apache Software Foundation

Apache Spark™ is a fast and general engine for large-scale data processing.

- What is Apache Spark?
- What do we do with Apache Spark?
- How do we execute programs on a Spark Cluster?

Latest News

Spark 2.1.2 released (Oct 09, 2017)

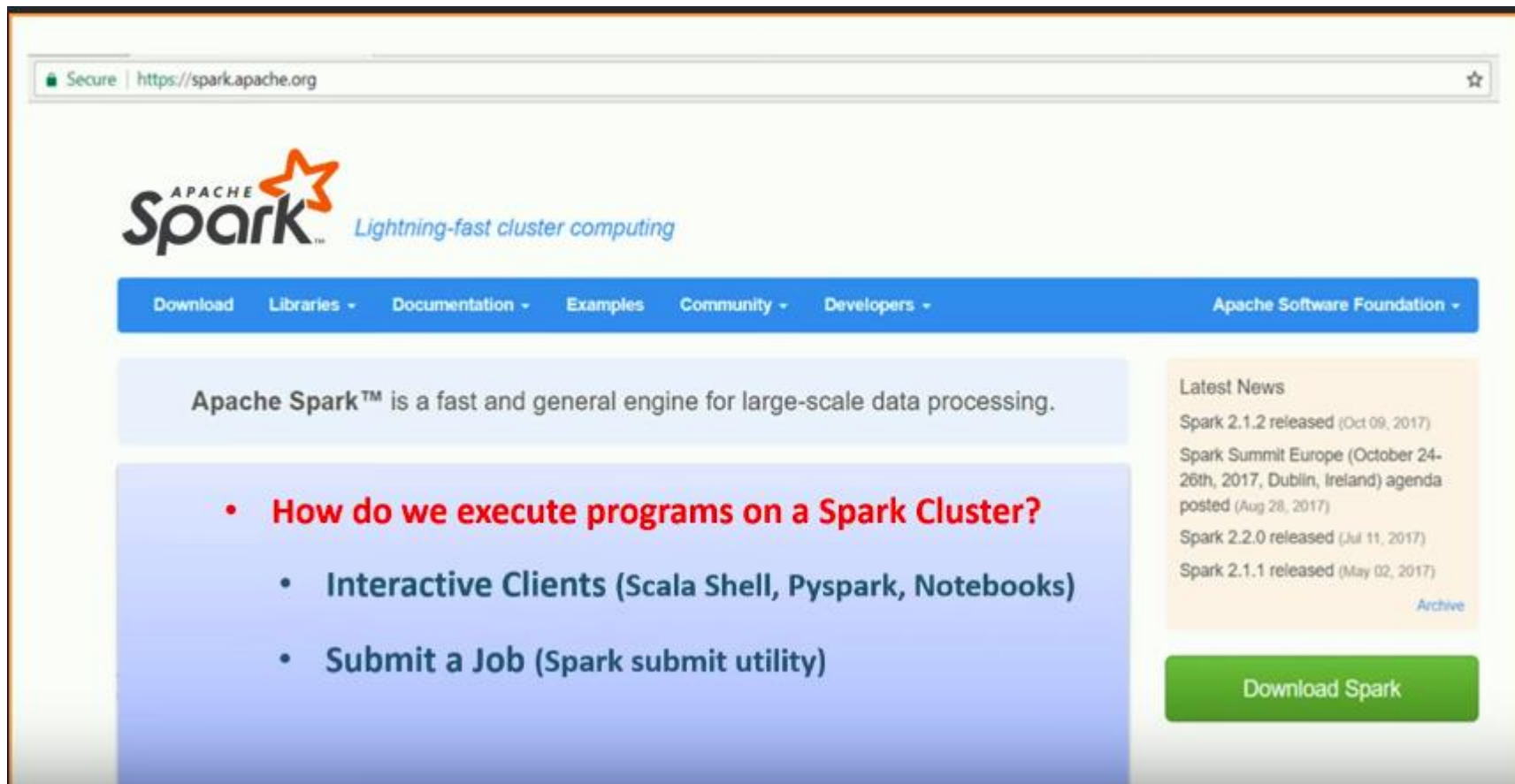
Spark Summit Europe (October 24-26th, 2017, Dublin, Ireland) agenda posted (Aug 28, 2017)

Spark 2.2.0 released (Jul 11, 2017)

Spark 2.1.1 released (May 02, 2017)

Archive

Download Spark



The screenshot shows the Apache Spark website at <https://spark.apache.org>. The page features the Apache Spark logo with the tagline "Lightning-fast cluster computing". A blue navigation bar contains links for Download, Libraries, Documentation, Examples, Community, Developers, and the Apache Software Foundation. Below the navigation bar, a light blue box states: "Apache Spark™ is a fast and general engine for large-scale data processing." A large blue box contains a red bullet point: "• How do we execute programs on a Spark Cluster?". Below this, two dark blue bullet points are listed: "• Interactive Clients (Scala Shell, Pyspark, Notebooks)" and "• Submit a Job (Spark submit utility)". On the right side, a "Latest News" section lists recent releases and events, including "Spark 2.1.2 released (Oct 09, 2017)", "Spark Summit Europe (October 24-26th, 2017, Dublin, Ireland) agenda posted (Aug 28, 2017)", "Spark 2.2.0 released (Jul 11, 2017)", and "Spark 2.1.1 released (May 02, 2017)". A green "Download Spark" button is located at the bottom right of the page.

Secure | <https://spark.apache.org>

APACHE Spark™ Lightning-fast cluster computing

Download Libraries - Documentation - Examples Community - Developers - Apache Software Foundation -

Apache Spark™ is a fast and general engine for large-scale data processing.

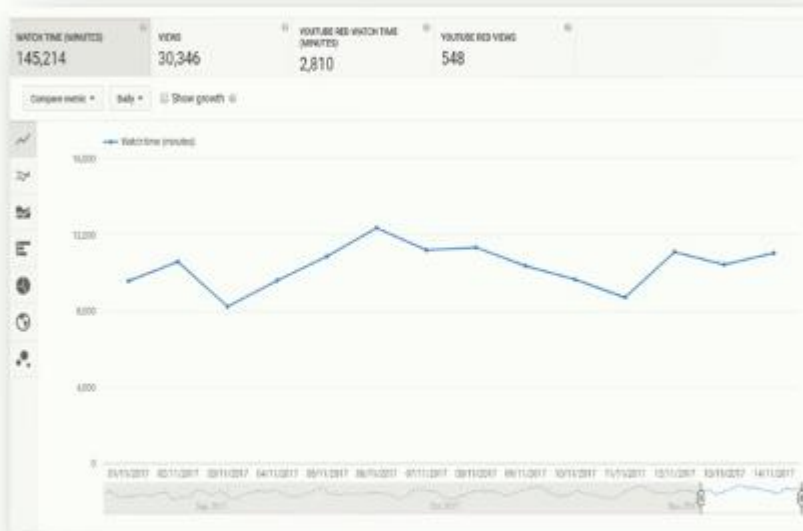
- **How do we execute programs on a Spark Cluster?**
 - Interactive Clients (Scala Shell, Pyspark, Notebooks)
 - Submit a Job (Spark submit utility)

Latest News

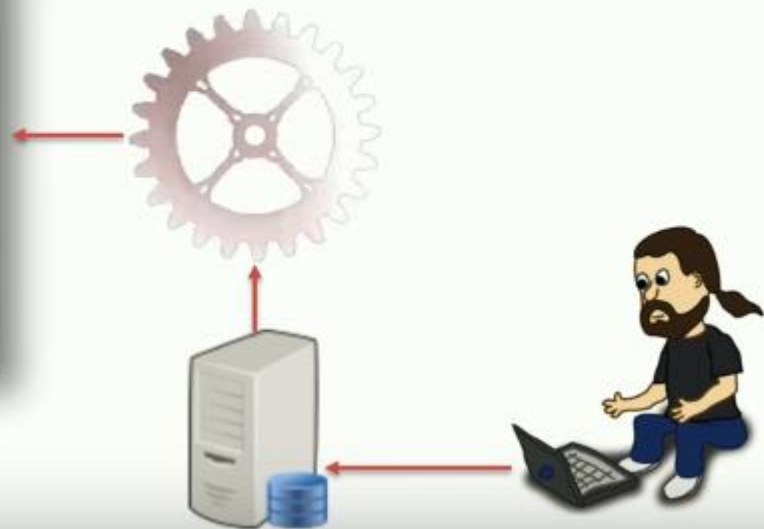
- Spark 2.1.2 released (Oct 09, 2017)
- Spark Summit Europe (October 24-26th, 2017, Dublin, Ireland) agenda posted (Aug 28, 2017)
- Spark 2.2.0 released (Jul 11, 2017)
- Spark 2.1.1 released (May 02, 2017)

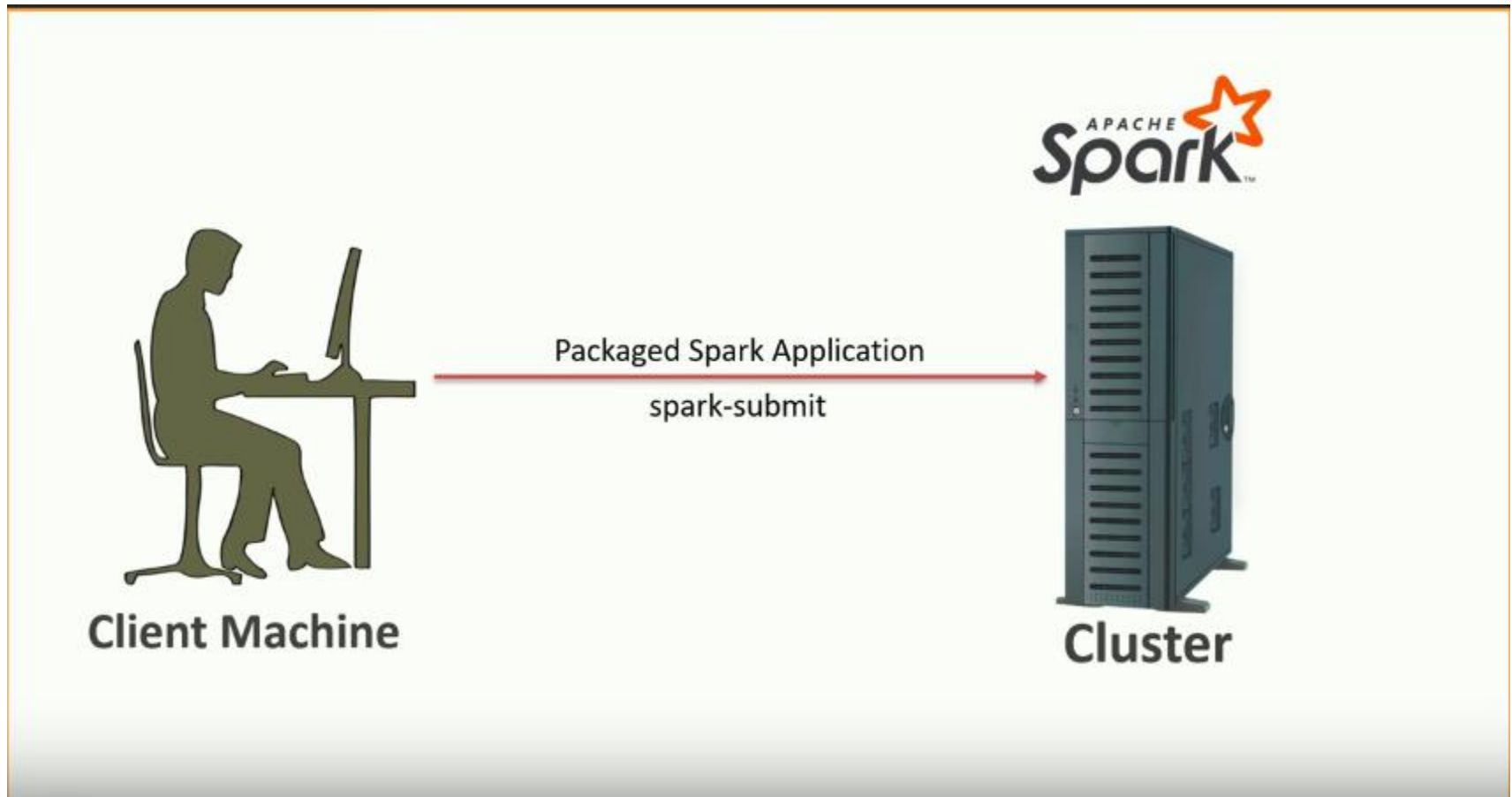
[Archive](#)

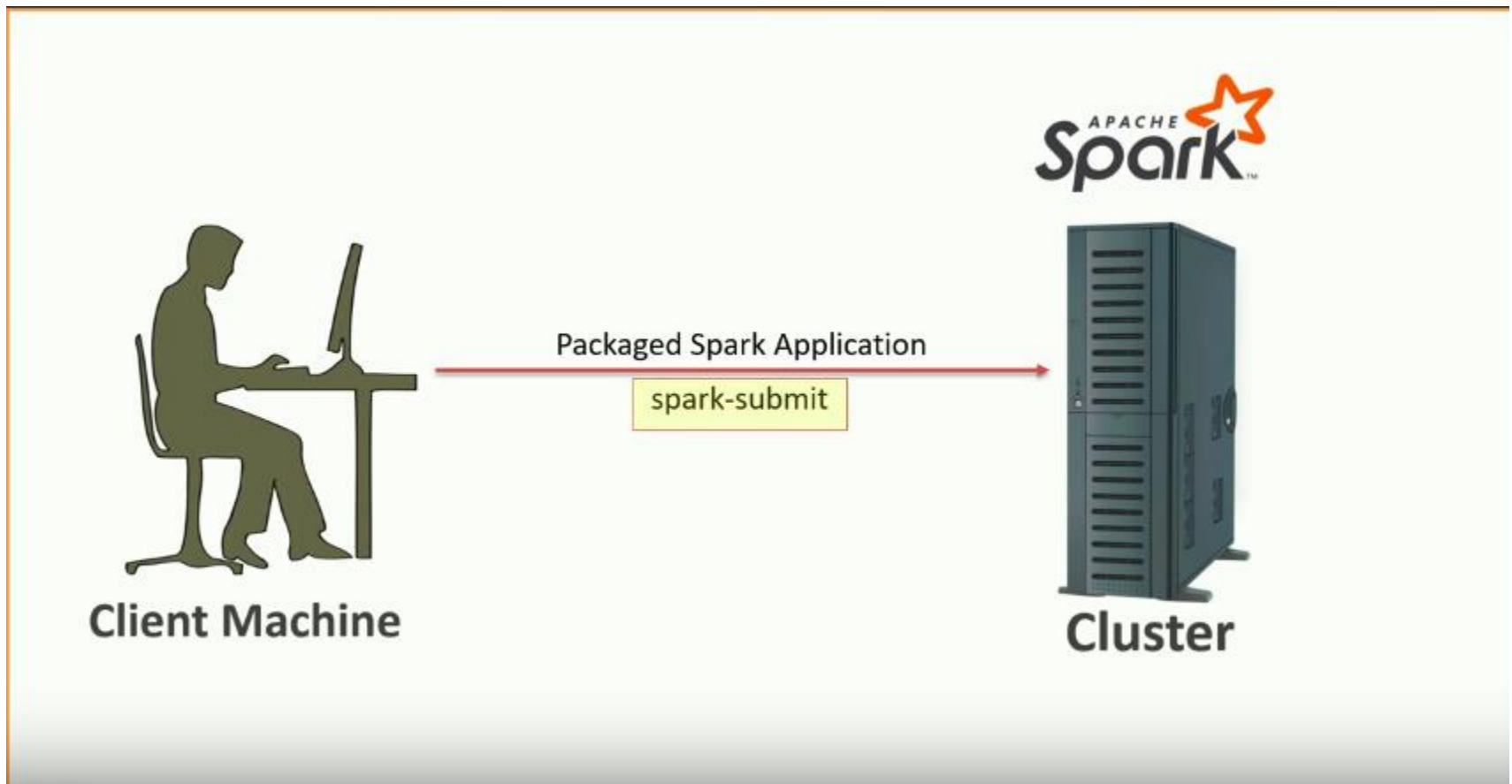
Download Spark

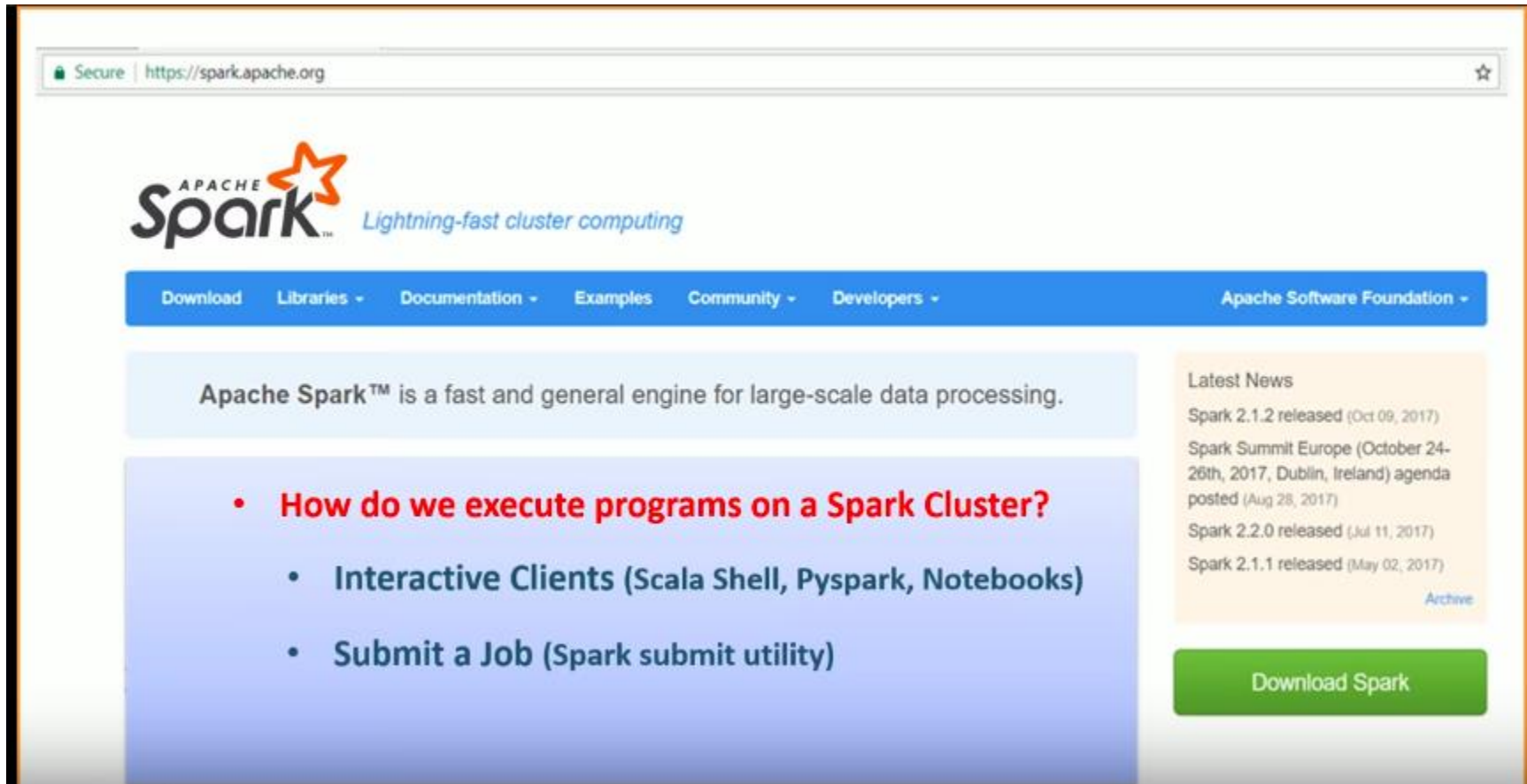


YouTube daily watch time in minutes









The screenshot shows the Apache Spark website. The browser address bar displays 'Secure | https://spark.apache.org'. The Apache Spark logo is prominently displayed, featuring the word 'APACHE' in small letters above 'Spark' in a large, bold font, with an orange star icon to the right. Below the logo, the tagline 'Lightning-fast cluster computing' is visible. A blue navigation bar contains links for 'Download', 'Libraries -', 'Documentation -', 'Examples', 'Community -', 'Developers -', and 'Apache Software Foundation -'. The main content area has a light blue background with the text 'Apache Spark™ is a fast and general engine for large-scale data processing.' Below this, a list of bullet points is shown: '• How do we execute programs on a Spark Cluster?' (in red), '• Interactive Clients (Scala Shell, Pyspark, Notebooks)', and '• Submit a Job (Spark submit utility)'. To the right, a 'Latest News' section lists recent releases: 'Spark 2.1.2 released (Oct 09, 2017)', 'Spark Summit Europe (October 24-26th, 2017, Dublin, Ireland) agenda posted (Aug 28, 2017)', 'Spark 2.2.0 released (Jul 11, 2017)', and 'Spark 2.1.1 released (May 02, 2017)', with an 'Archive' link below. A green 'Download Spark' button is located at the bottom right of the main content area.

Secure | https://spark.apache.org

APACHE Spark™ Lightning-fast cluster computing

Download Libraries - Documentation - Examples Community - Developers - Apache Software Foundation -

Apache Spark™ is a fast and general engine for large-scale data processing.

- **How do we execute programs on a Spark Cluster?**
 - Interactive Clients (Scala Shell, Pyspark, Notebooks)
 - Submit a Job (Spark submit utility)

Latest News

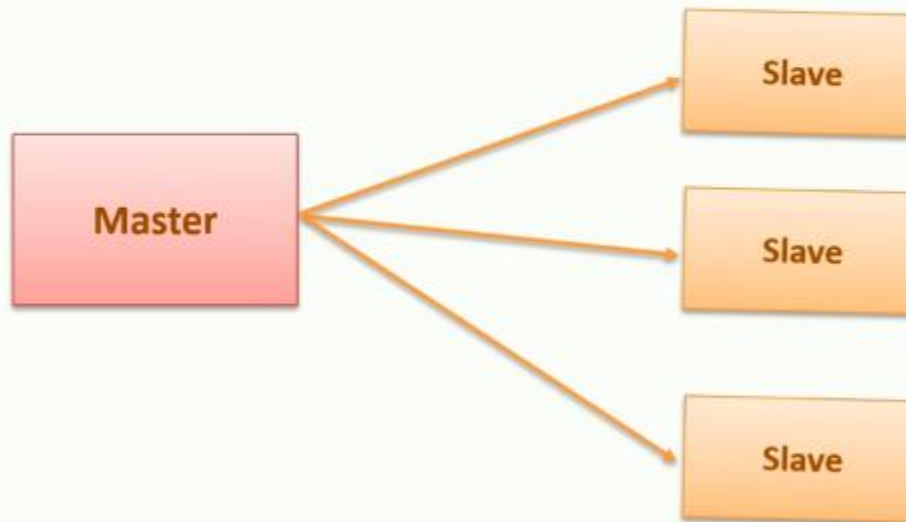
- Spark 2.1.2 released (Oct 09, 2017)
- Spark Summit Europe (October 24-26th, 2017, Dublin, Ireland) agenda posted (Aug 28, 2017)
- Spark 2.2.0 released (Jul 11, 2017)
- Spark 2.1.1 released (May 02, 2017)

[Archive](#)

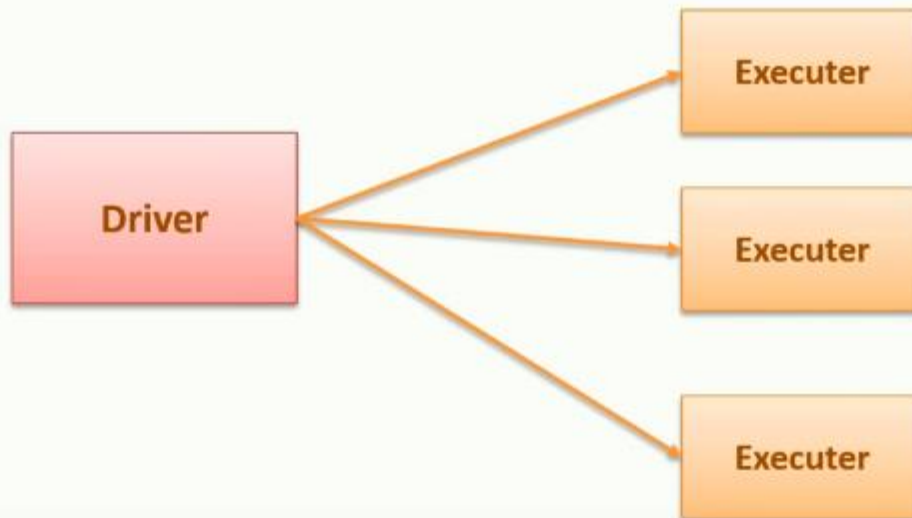
Download Spark

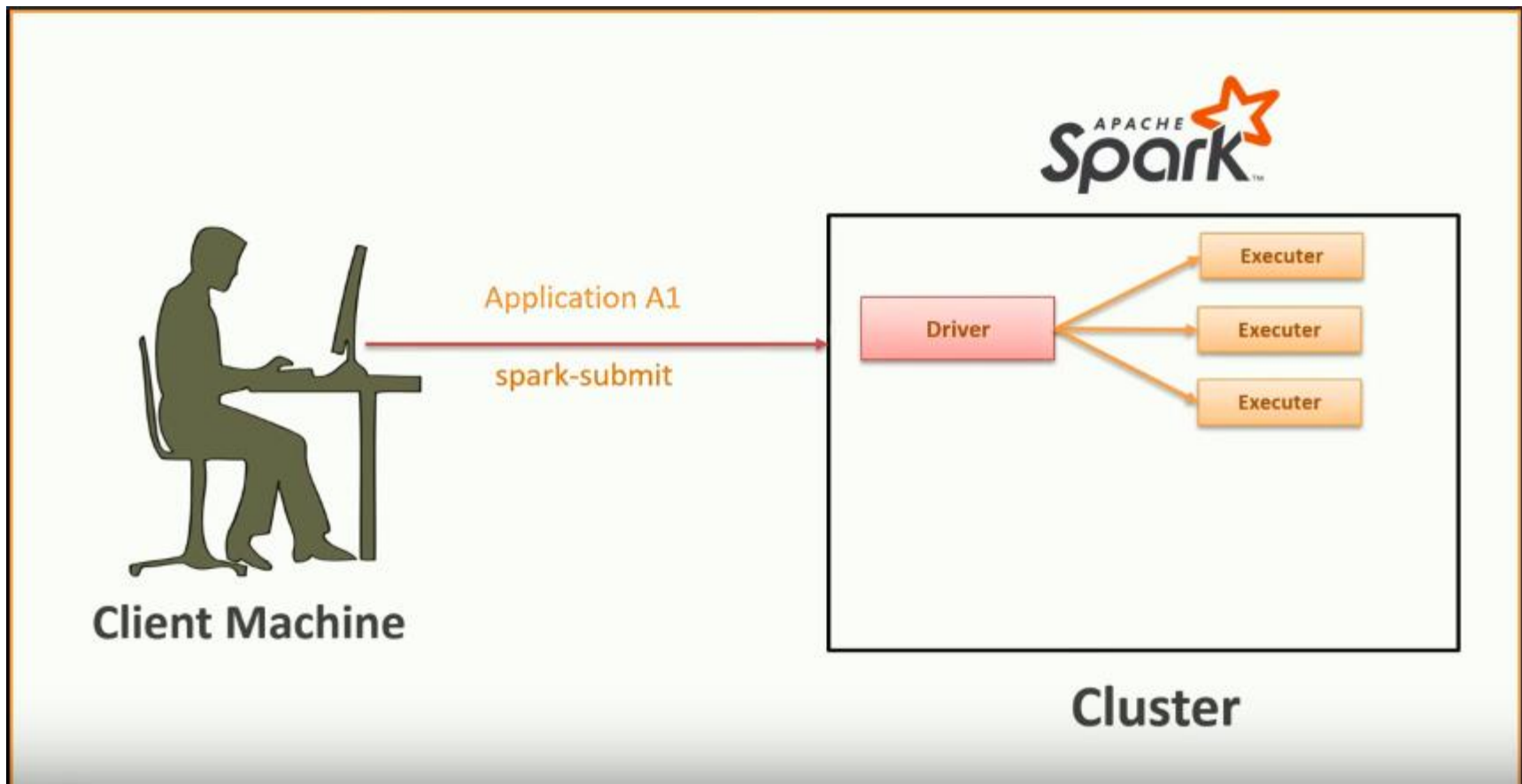
How does the Spark execute our programs on a cluster?

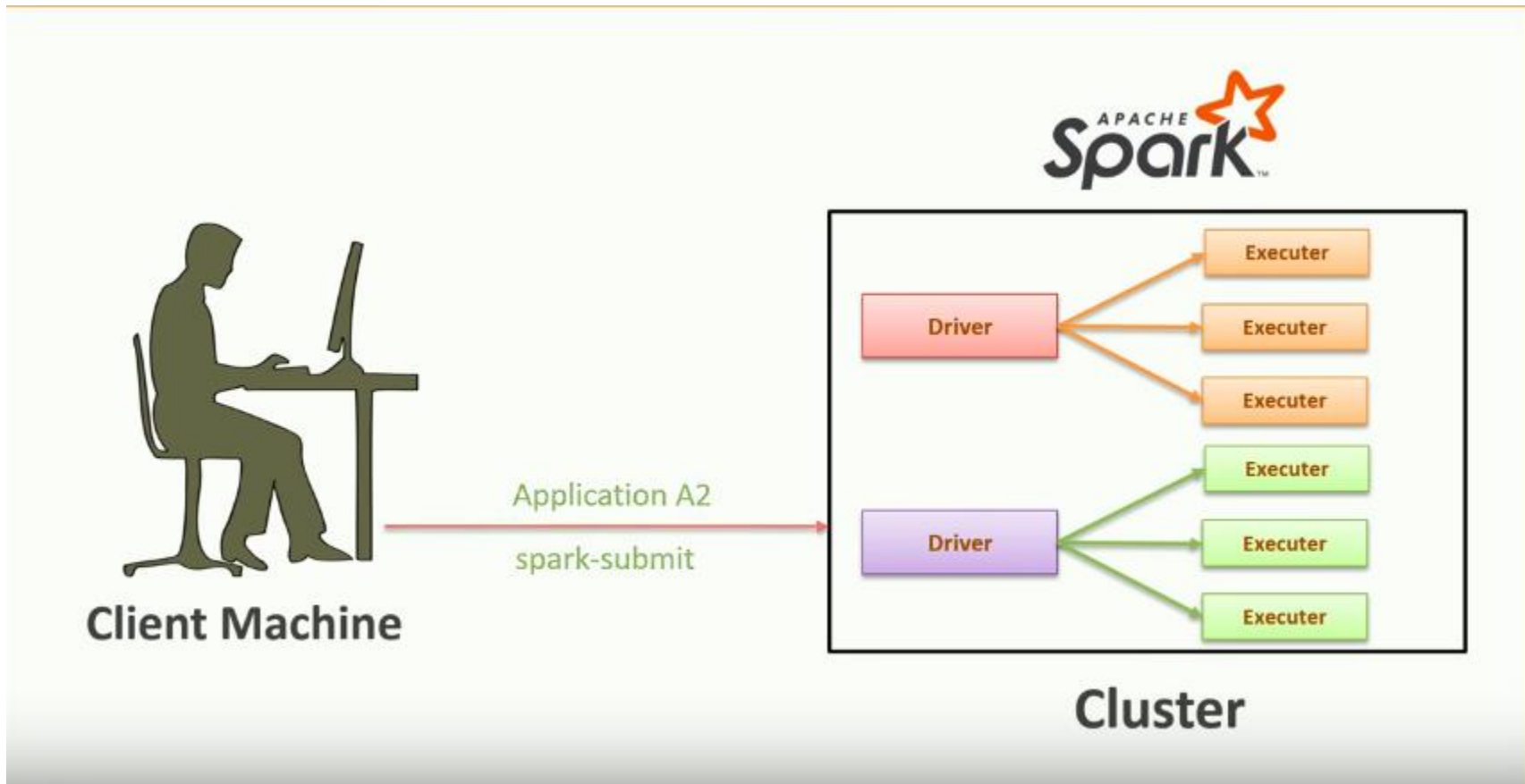
How does the Spark execute our programs on a cluster?

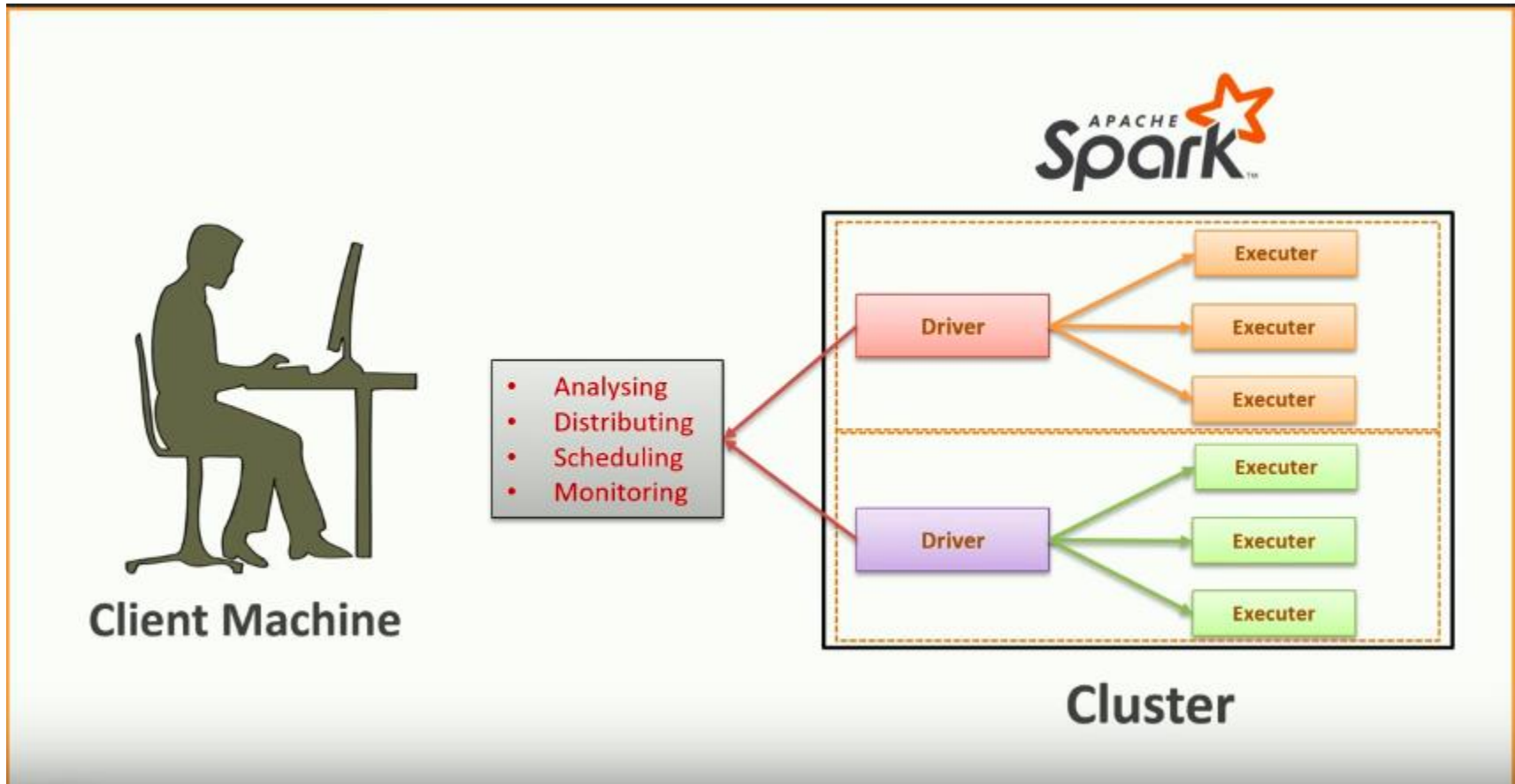


How does the Spark execute our programs on a cluster?









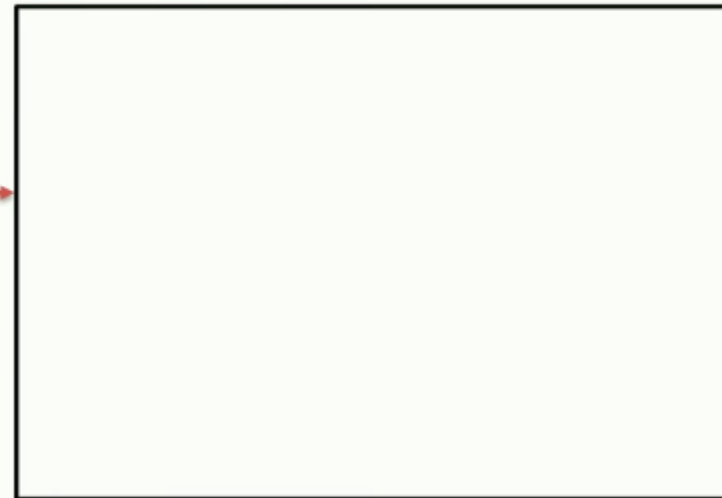
Who executes Where?

Who executes Where?



Client Machine

Application



Cluster

Who executes Where?



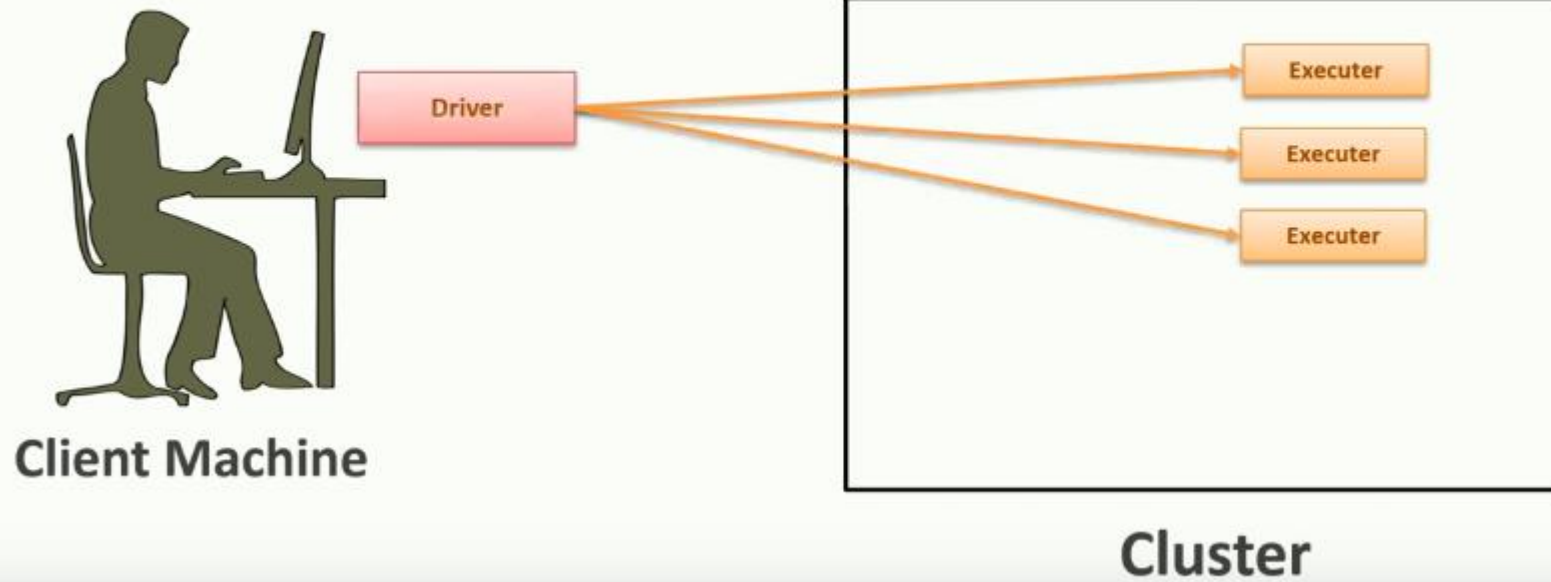
Client Machine

Application



Cluster

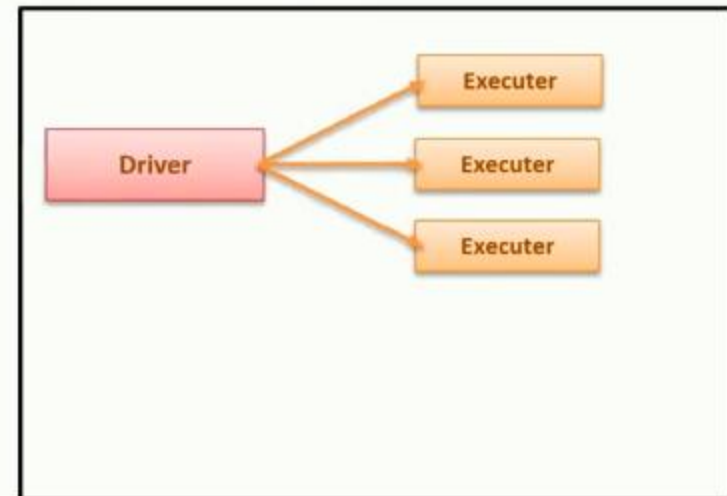
Who executes Where?



Who executes Where?



Client Machine



Cluster

Who executes Where?

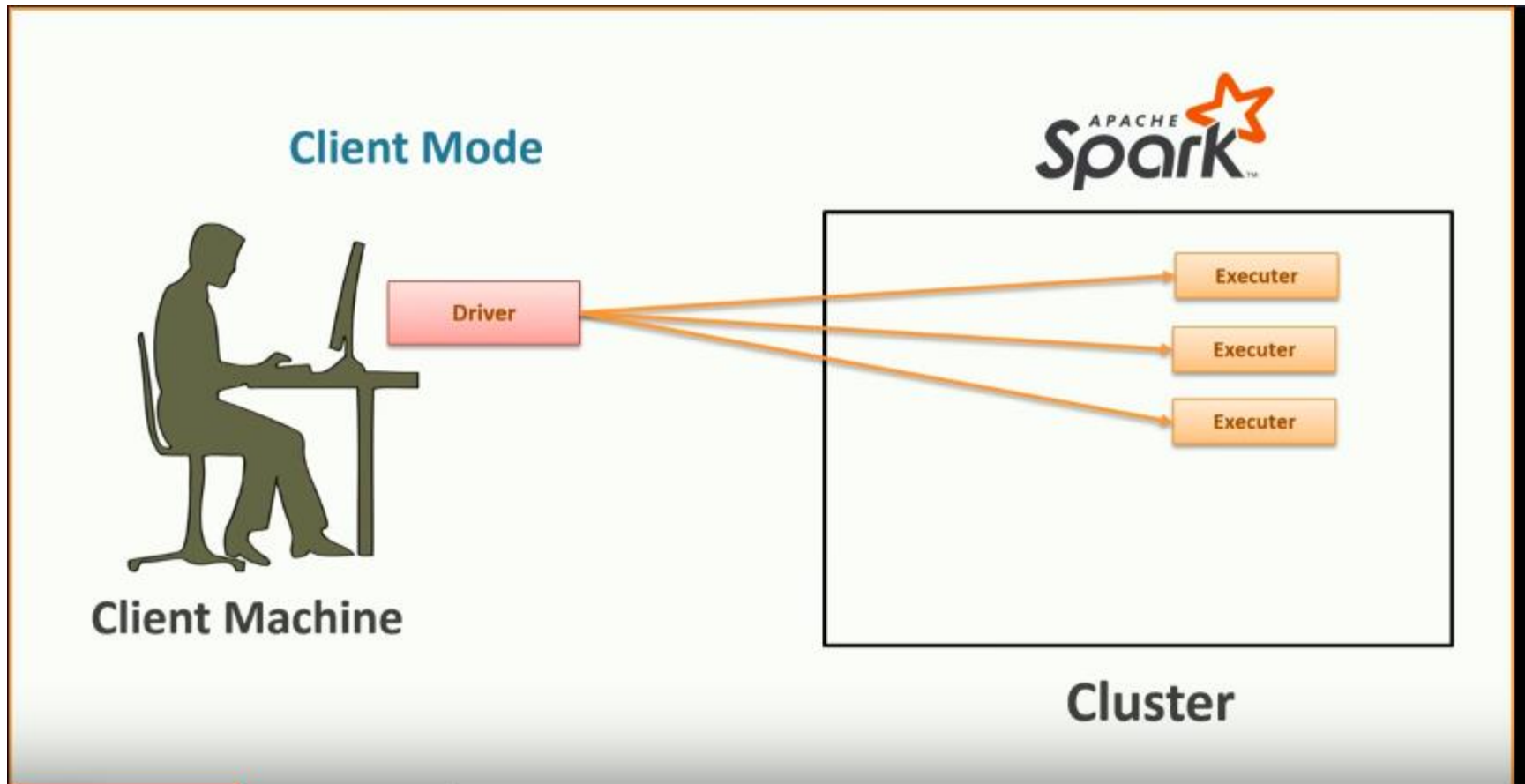


Client Machine

1. Client Mode
2. Cluster Mode



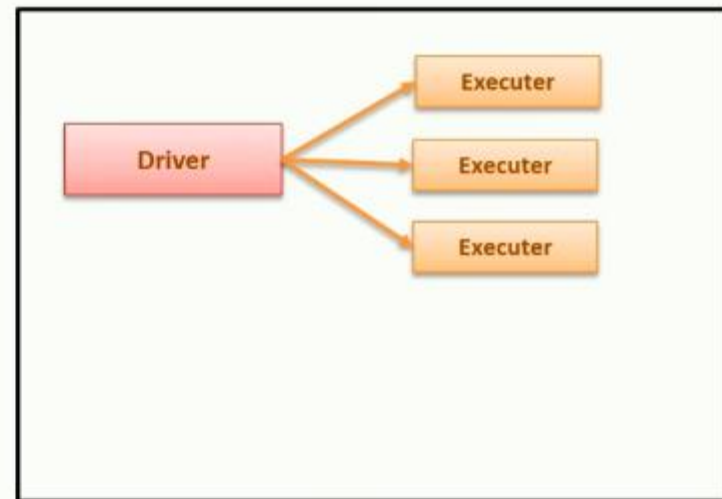
Cluster



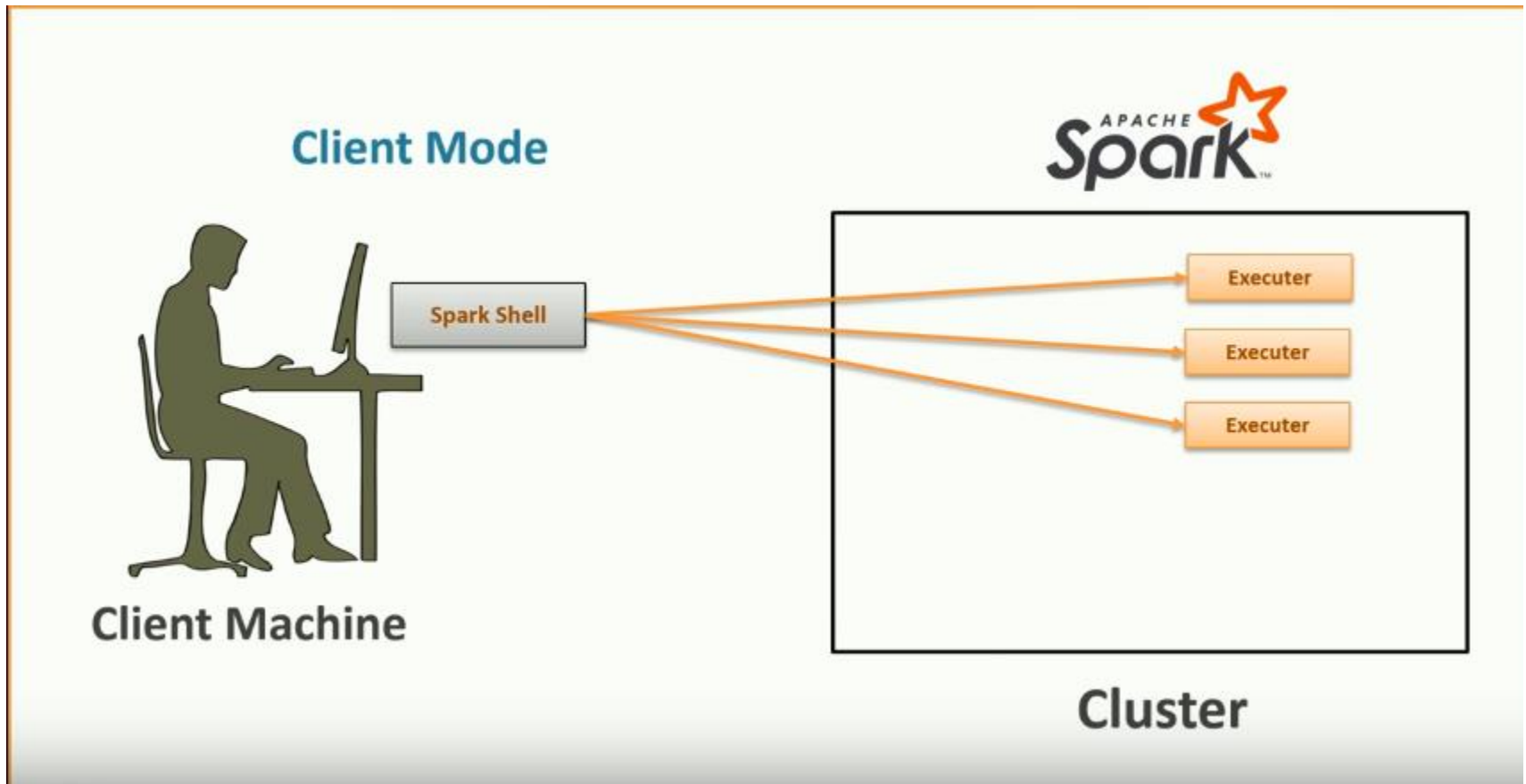
Cluster Mode



Client Machine



Cluster



How does the Spark execute our programs on a cluster?

How does the Spark execute our programs on a cluster?

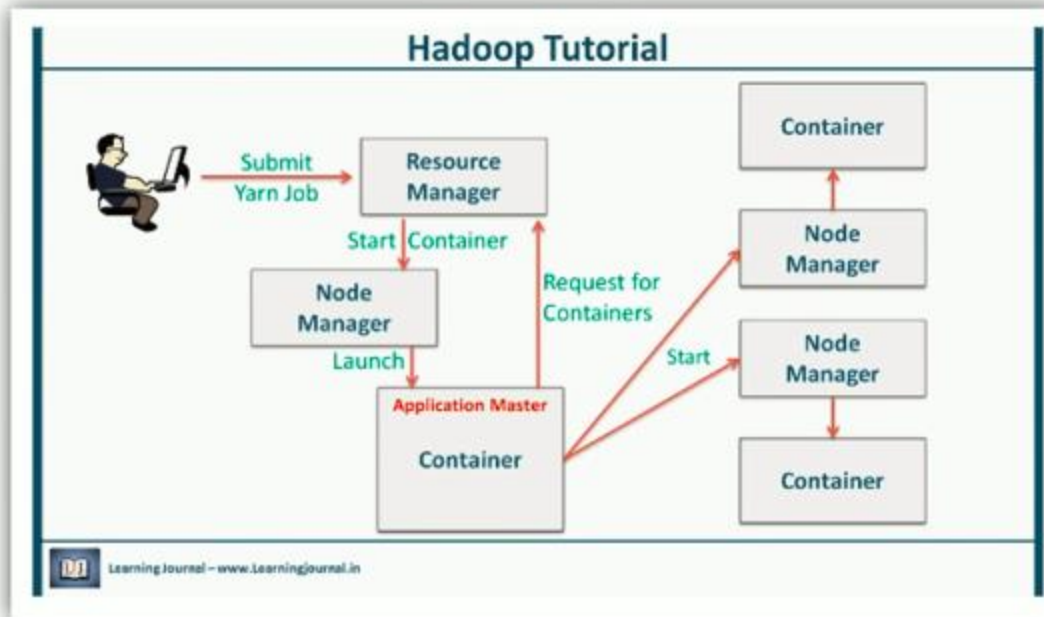
- Spark creates one driver and a bunch of executors for each application.
- Spark offers two deployment modes for an application.
 - Client Mode - Driver on client machine and Executors on cluster
 - Cluster Mode - Driver and executors on cluster

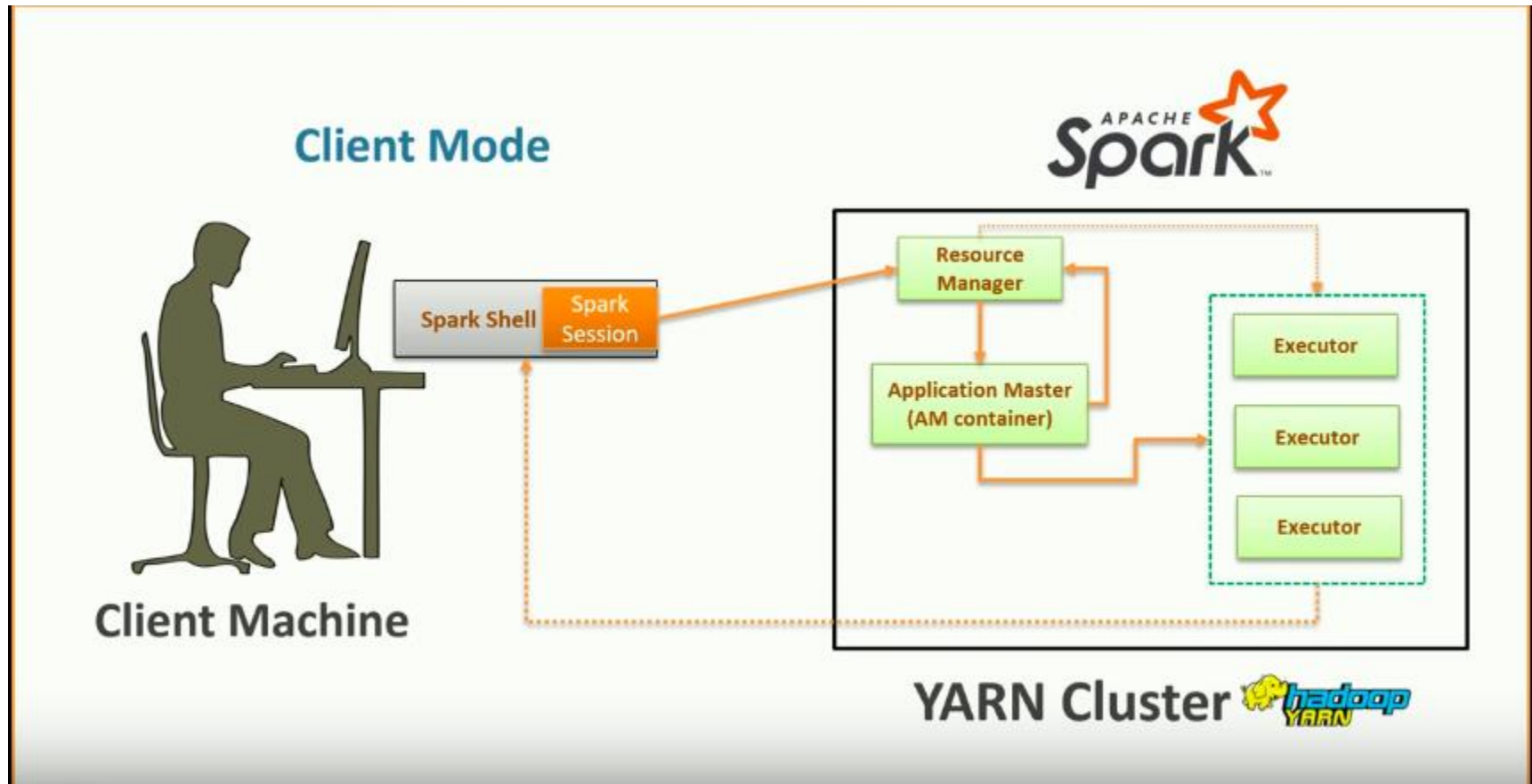
Who controls the cluster?

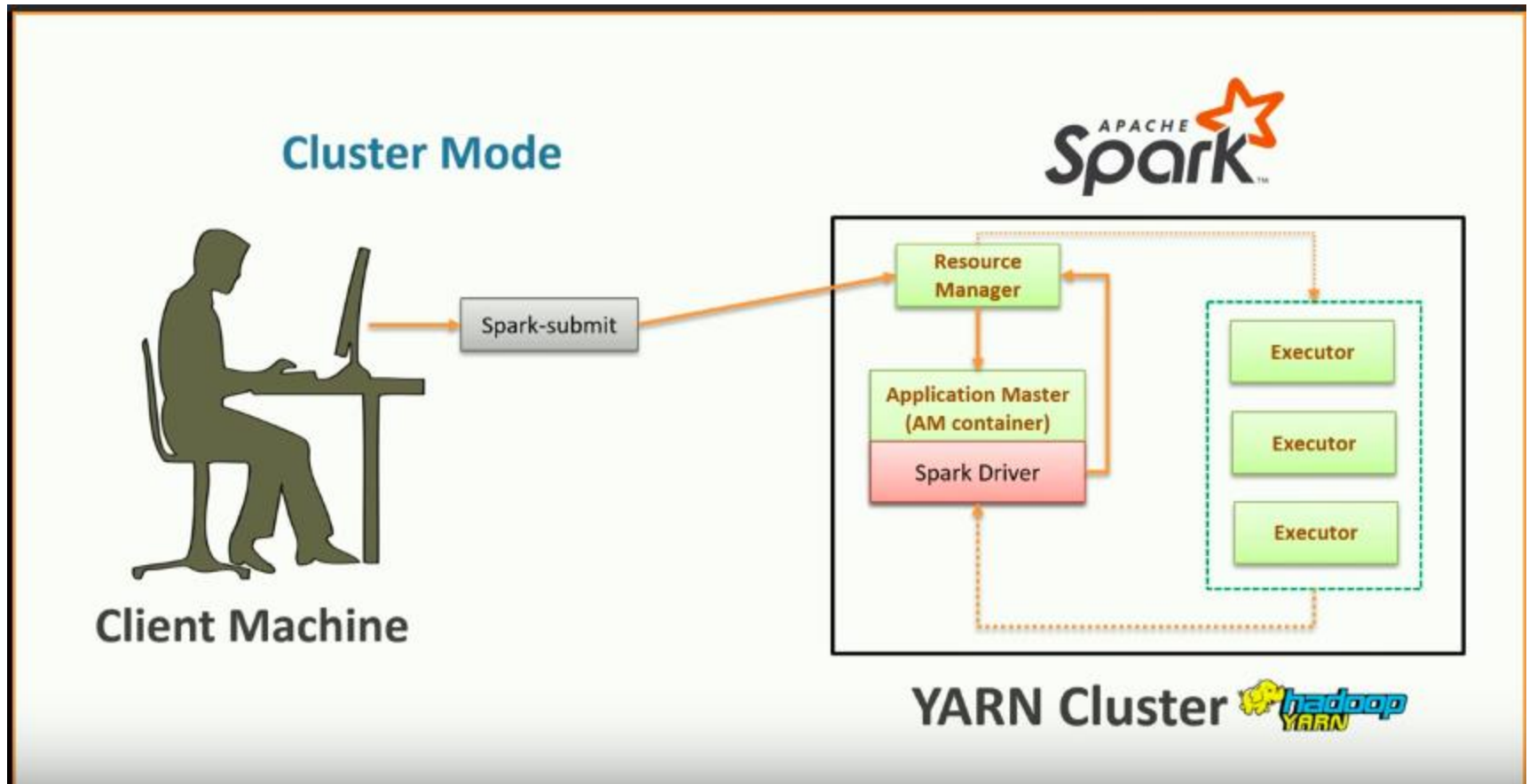
How Spark gets the resources for the driver and the executors?

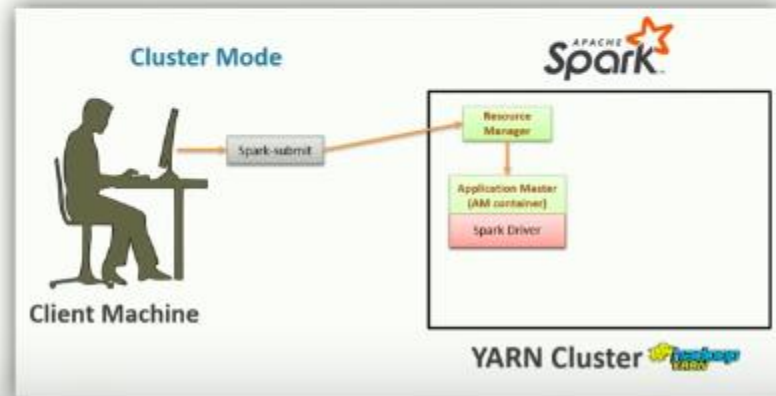
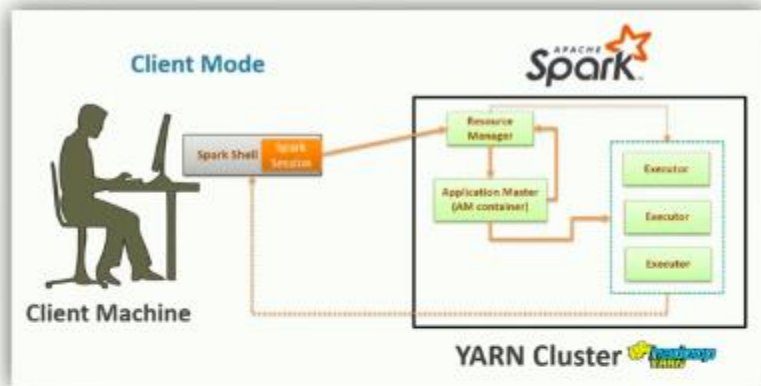
1. Apache YARN
2. Apache Mesos
3. Kubernetes
4. Standalone







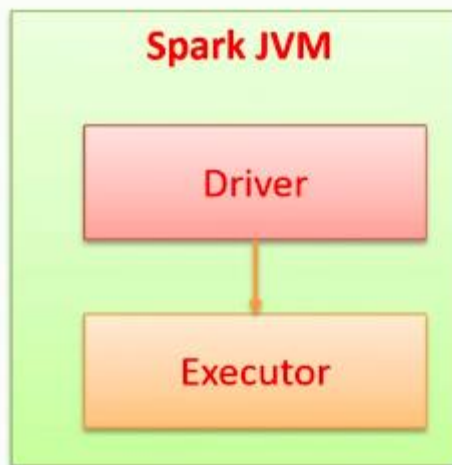




Local Mode



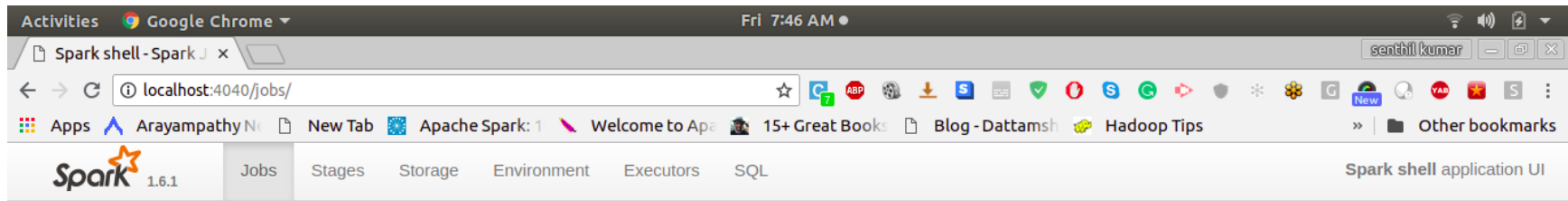
Local Machine



Demo

1. Driver
2. Executors
3. Client mode
4. Cluster mode
5. Local mode.

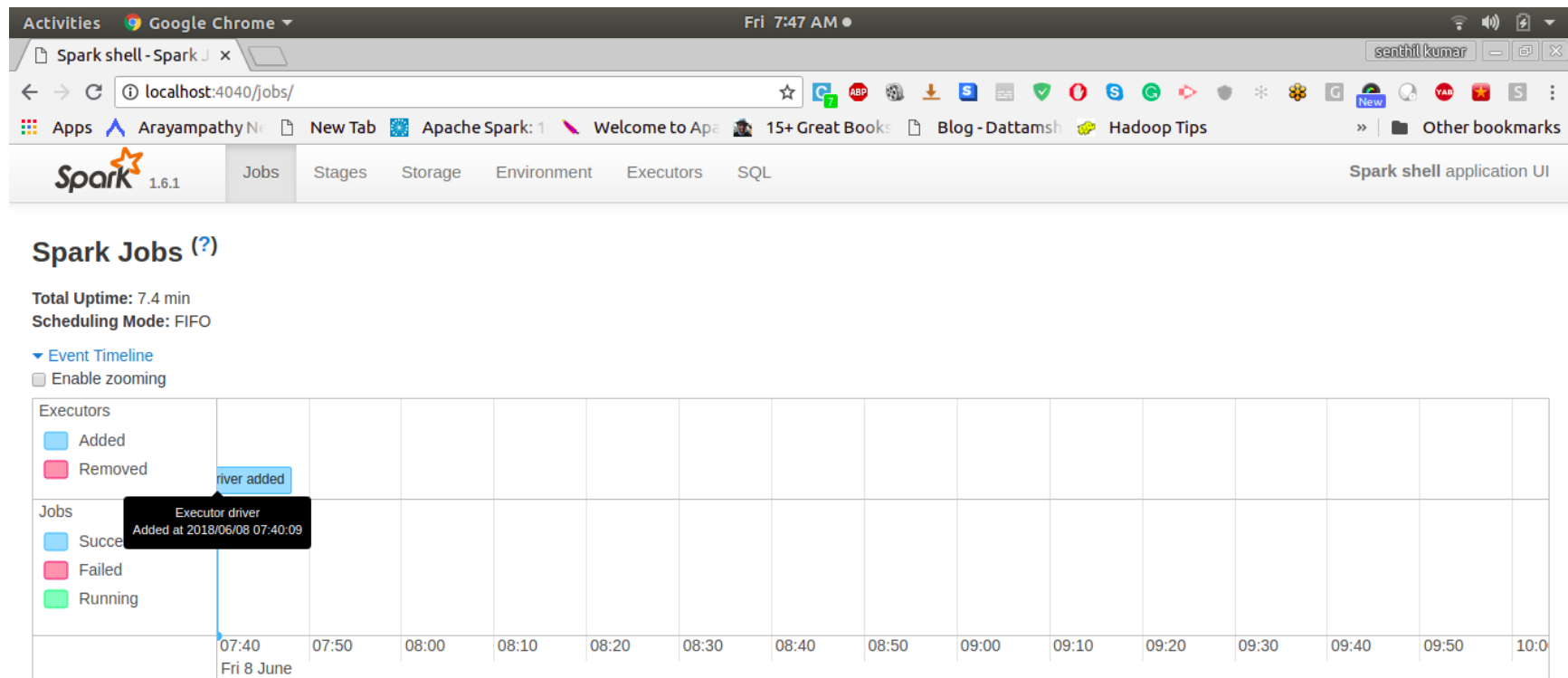
1. An interactive method using the spark shell.
2. Fire an application using Spark-Submit utility.

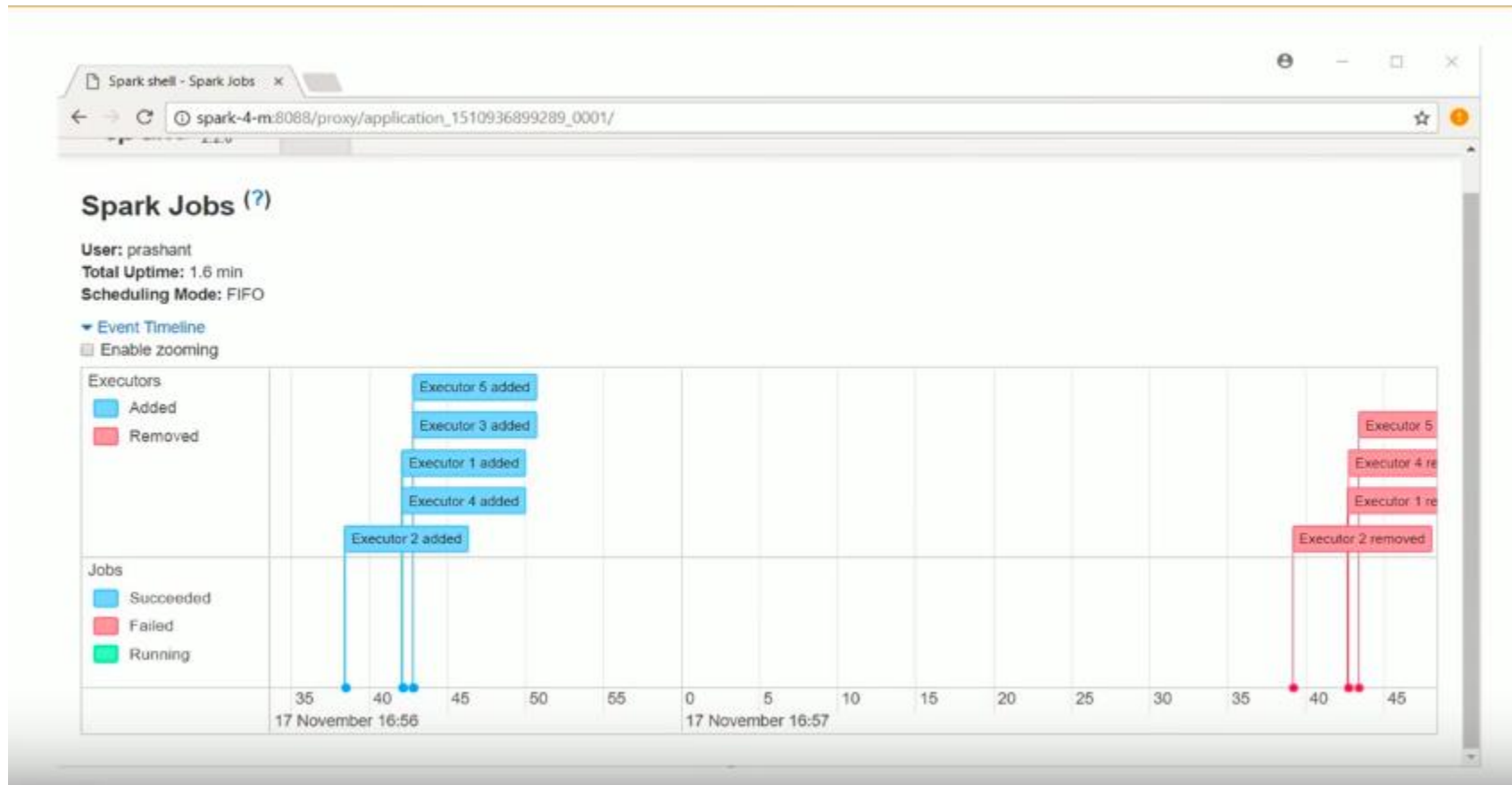


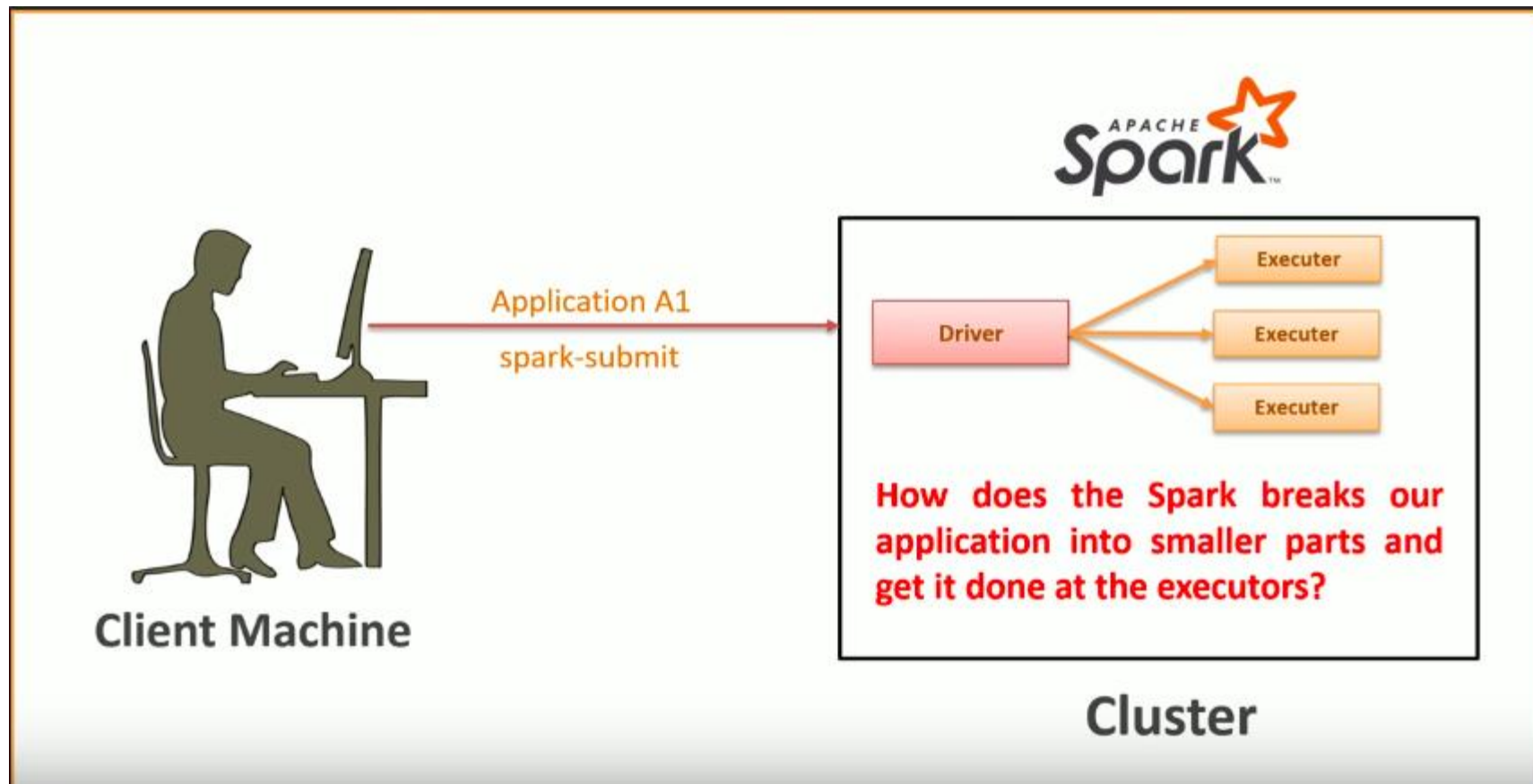
Spark Jobs (?)

Total Uptime: 5.9 min
Scheduling Mode: FIFO

[▶ Event Timeline](#)







Thank you