Installing, configuring, and running Cassandra locally

Apache Cassandra:
Core Concepts, Skills, and Tools

Leo Schuman, Joe Chu

Oct 20, 2014

Learning Objectives

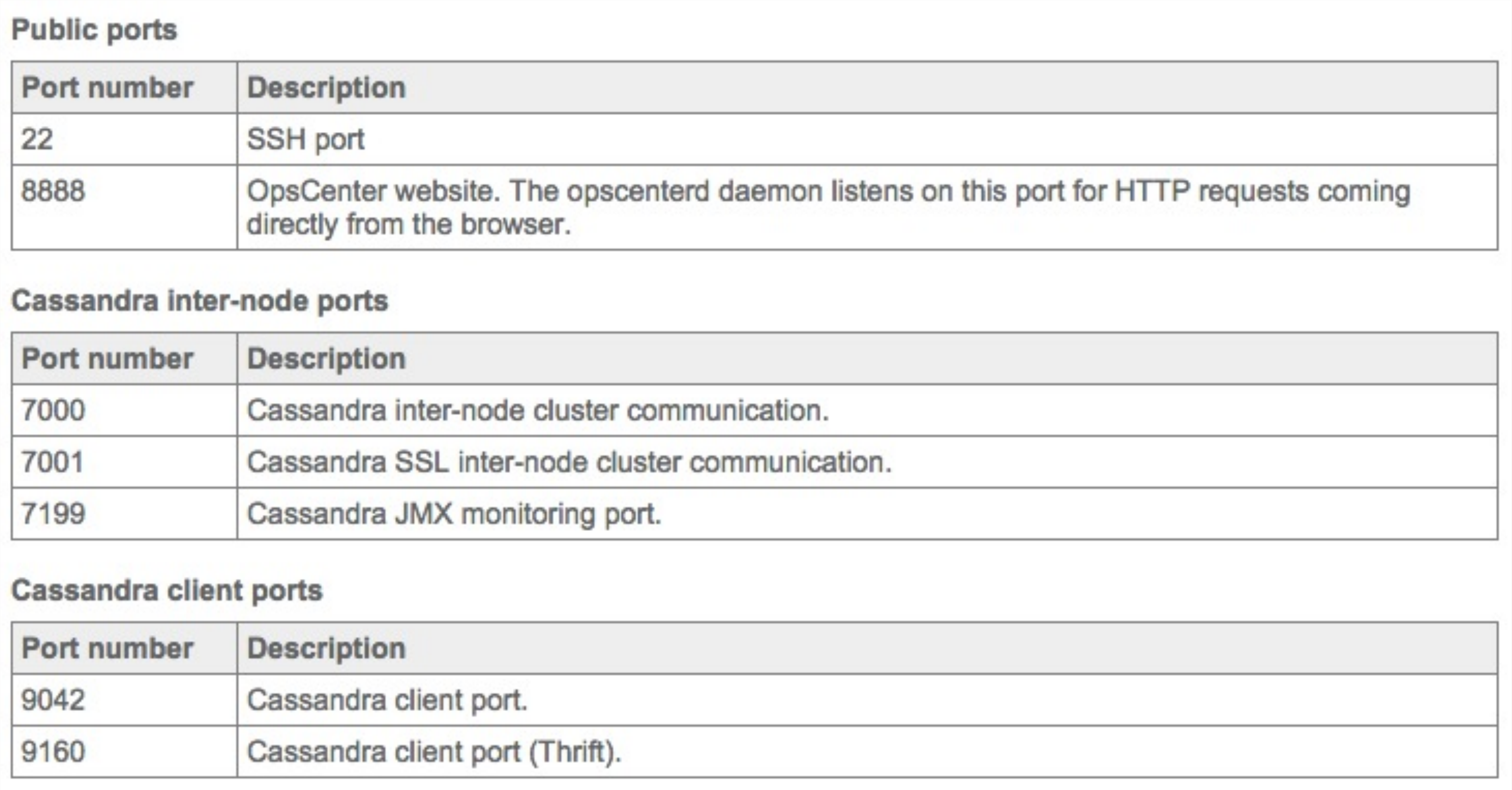
- **Prepare the operating system**
- Select and install a Cassandra distribution
- Configure Cassandra for a single node
- Start and stop a Cassandra instance

How is the OS prepared for Cassandra and OpsCenter?

- Install latest Java 7 release
  - Oracle JDK 1.7+ required for Cassandra 2.0+
  - 64 bit Oracle Java 7 preferred
- Configure JAVA_HOME
  - JAVA_HOME=/usr/local/java/*jdk1.7.0_xx*
- Install Java Native Access (JNA) libraries (prior to C* 2.1)
  - required for production systems
- Synchronize clocks on each node system
  - use NTP or similar tool
- Disable swap
  - *sudo swapoff –all*
  - better Cassandra be killed for lack of memory than bogged by JVM swap

How must ports be configured?

- Verify these ports are open and available



Public ports

Port number	Description
22	SSH port
8888	OpsCenter website. The opscenterd daemon listens on this port for HTTP requests coming directly from the browser.

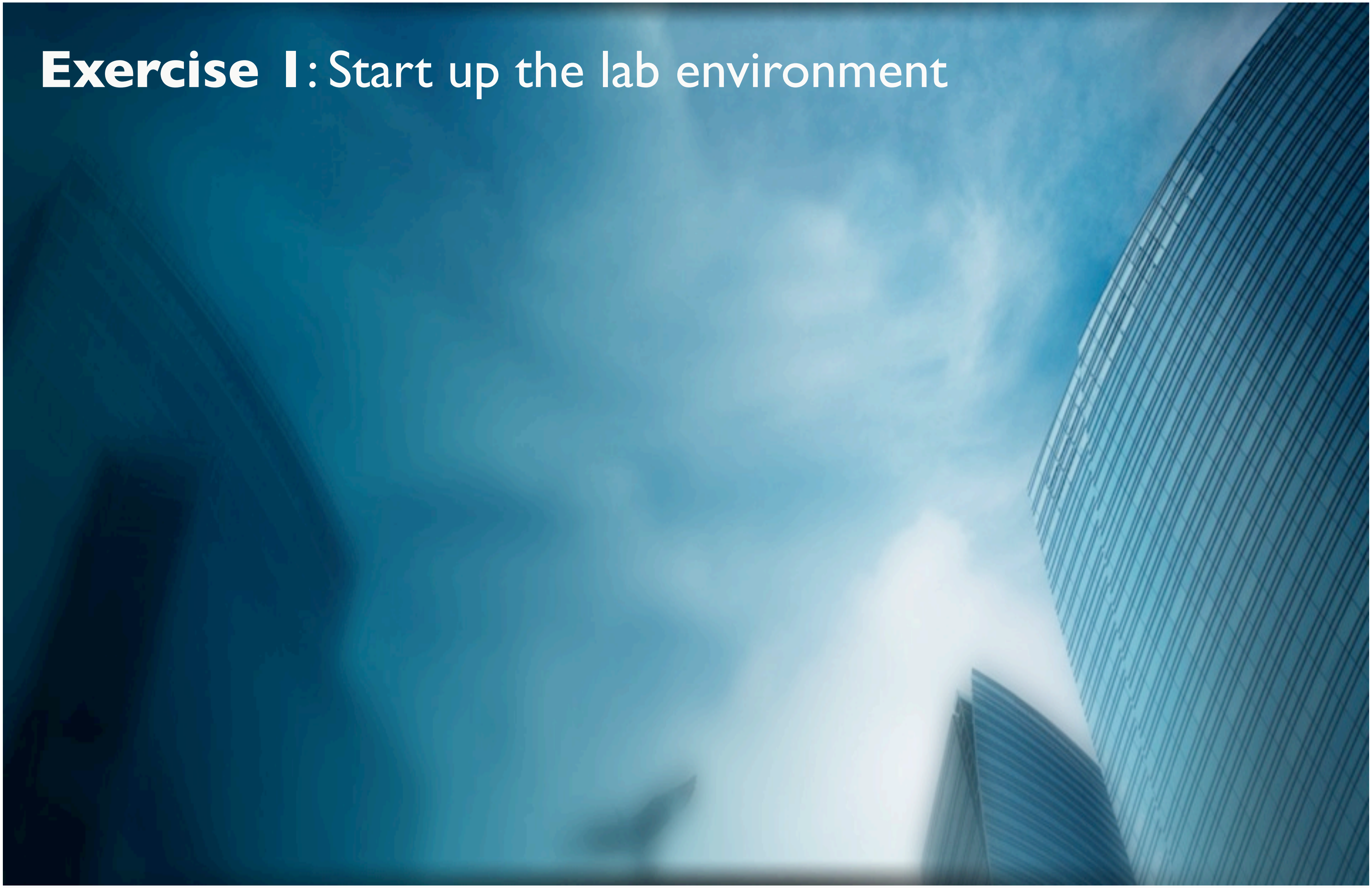
Cassandra inter-node ports

Port number	Description
7000	Cassandra inter-node cluster communication.
7001	Cassandra SSL inter-node cluster communication.
7199	Cassandra JMX monitoring port.

Cassandra client ports

Port number	Description
9042	Cassandra client port.
9160	Cassandra client port (Thrift).

Exercise 1: Start up the lab environment

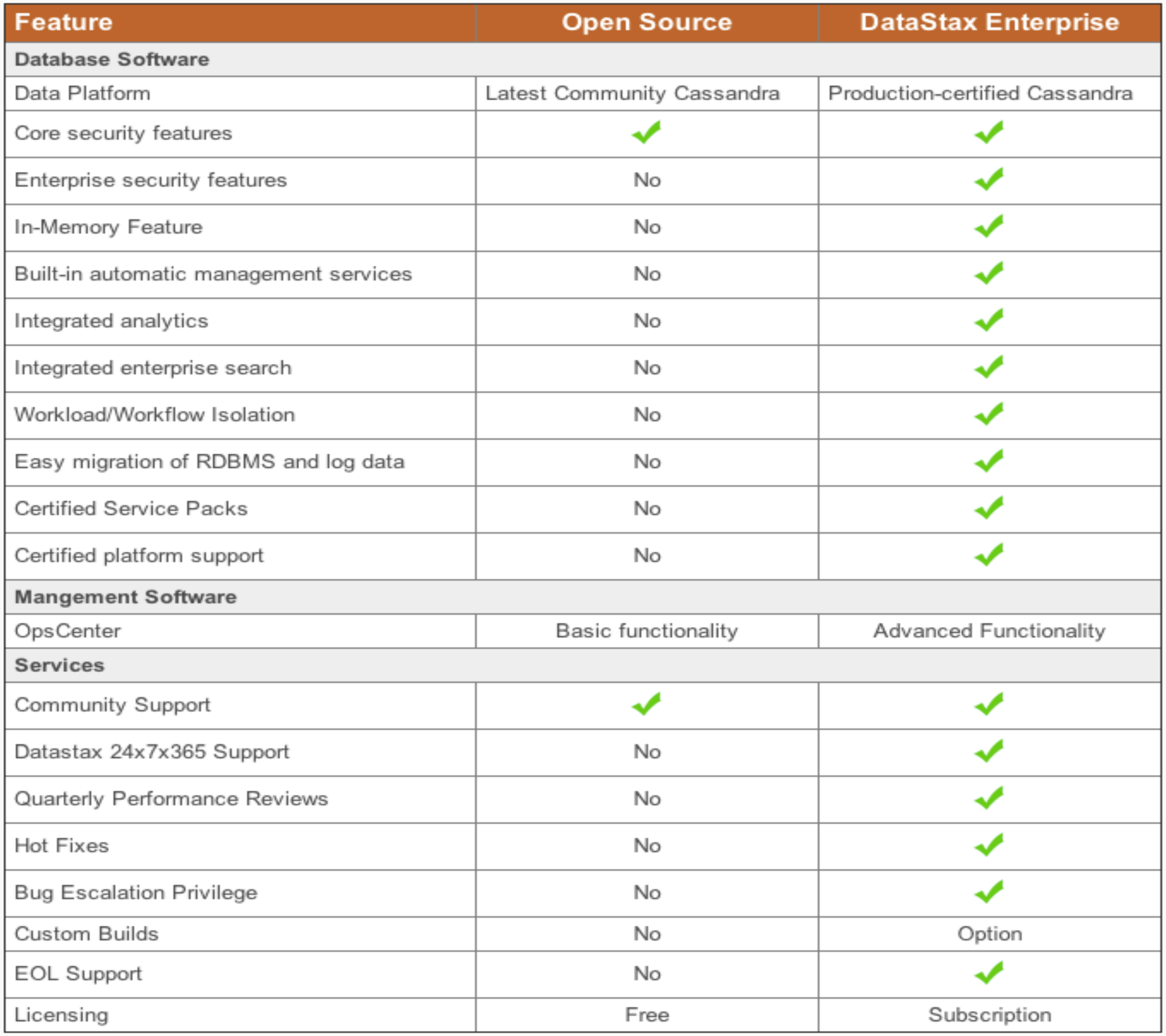


Learning Objectives

- Prepare the operating system
- **Select and install a Cassandra distribution**
- Configure Cassandra for a single node
- Start and stop a Cassandra instance

How do Cassandra, DSC, and DSE compare?

- Three distributions
  - DataStax Enterprise (DSE)
  - DataStax Community Edition (DSC)
  - Apache Cassandra



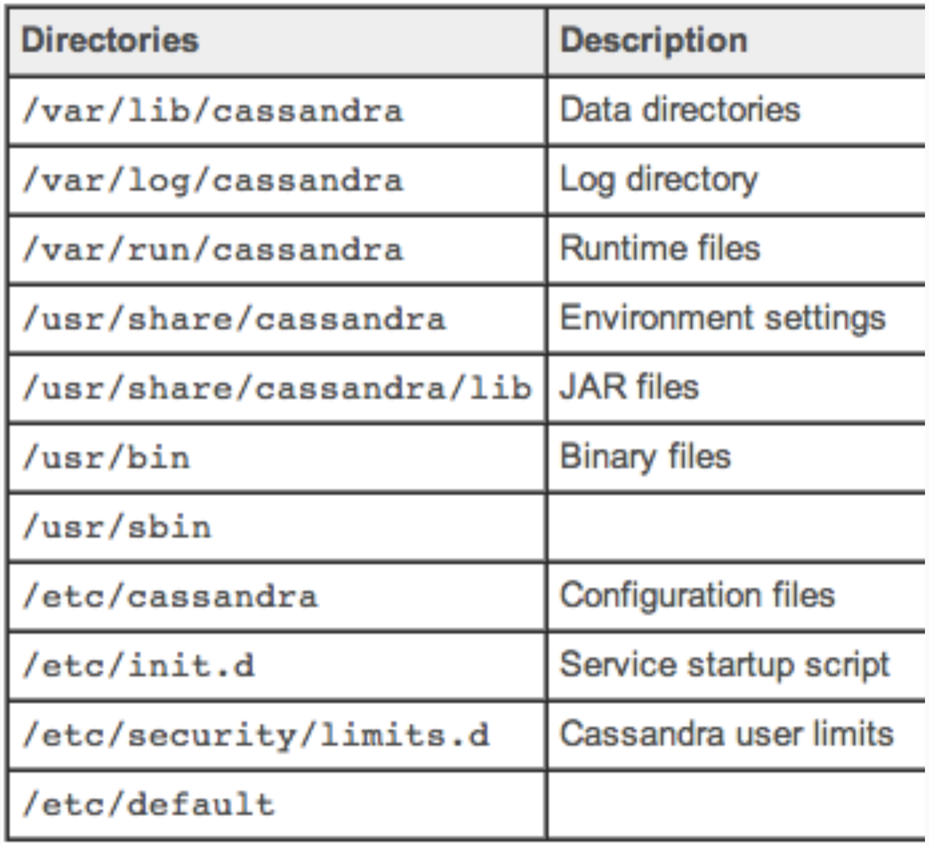
Feature	Open Source	DataStax Enterprise
Database Software		
Data Platform	Latest Community Cassandra	Production-certified Cassandra
Core security features	✓	✓
Enterprise security features	No	✓
In-Memory Feature	No	✓
Built-in automatic management services	No	✓
Integrated analytics	No	✓
Integrated enterprise search	No	✓
Workload/Workflow Isolation	No	✓
Easy migration of RDBMS and log data	No	✓
Certified Service Packs	No	✓
Certified platform support	No	✓
Mangement Software		
OpsCenter	Basic functionality	Advanced Functionality
Services		
Community Support	✓	✓
Datastax 24x7x365 Support	No	✓
Quarterly Performance Reviews	No	✓
Hot Fixes	No	✓
Bug Escalation Privilege	No	✓
Custom Builds	No	Option
EOL Support	No	✓
Licensing	Free	Subscription

Where can each distribution be found?

- DataStax Enterprise + OpsCenter, DevCenter, and Drivers
  - http://www.datastax.com/download
- DataStax Community Edition
  - http://planetcassandra.org/Download/DataStaxCommunityEdition
- Apache Cassandra
  - http://cassandra.apache.org/download/
  - https://github.com/apache/cassandra

Can Cassandra be installed by package?

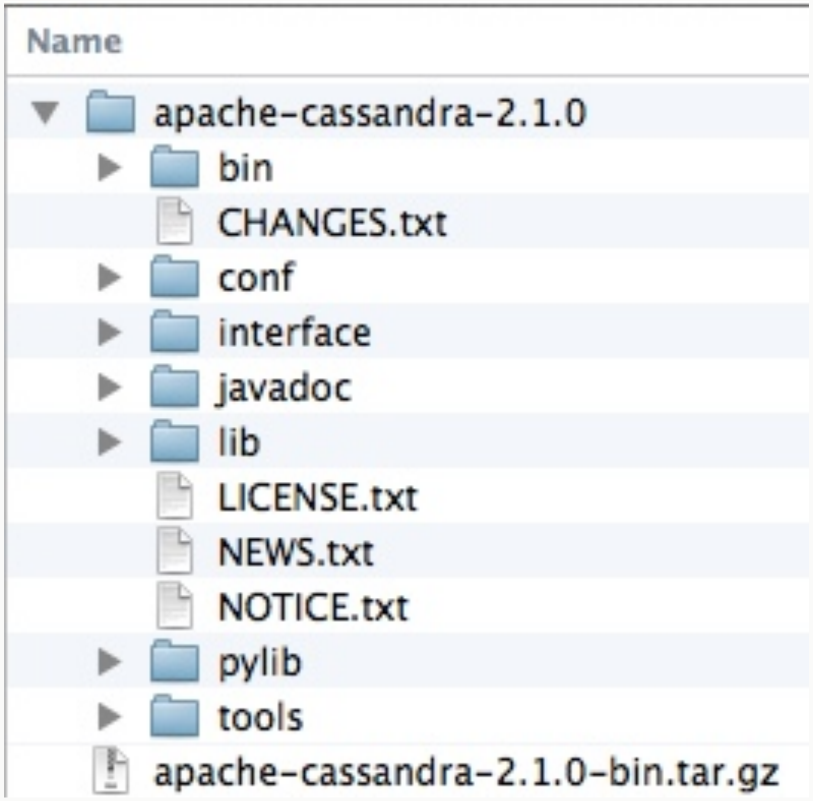
- Cassandra may be installed as a package
  - RPM on *nix using *yum*
  - DEB on *nix using *apt-get*
  - MSI on Windows
- Package installations define various folders for
  - data directories
  - log files
  - configuration files
  - binaries
  - security



Directories	Description
/var/lib/cassandra	Data directories
/var/log/cassandra	Log directory
/var/run/cassandra	Runtime files
/usr/share/cassandra	Environment settings
/usr/share/cassandra/lib	JAR files
/usr/bin	Binary files
/usr/sbin	
/etc/cassandra	Configuration files
/etc/init.d	Service startup script
/etc/security/limits.d	Cassandra user limits
/etc/default	

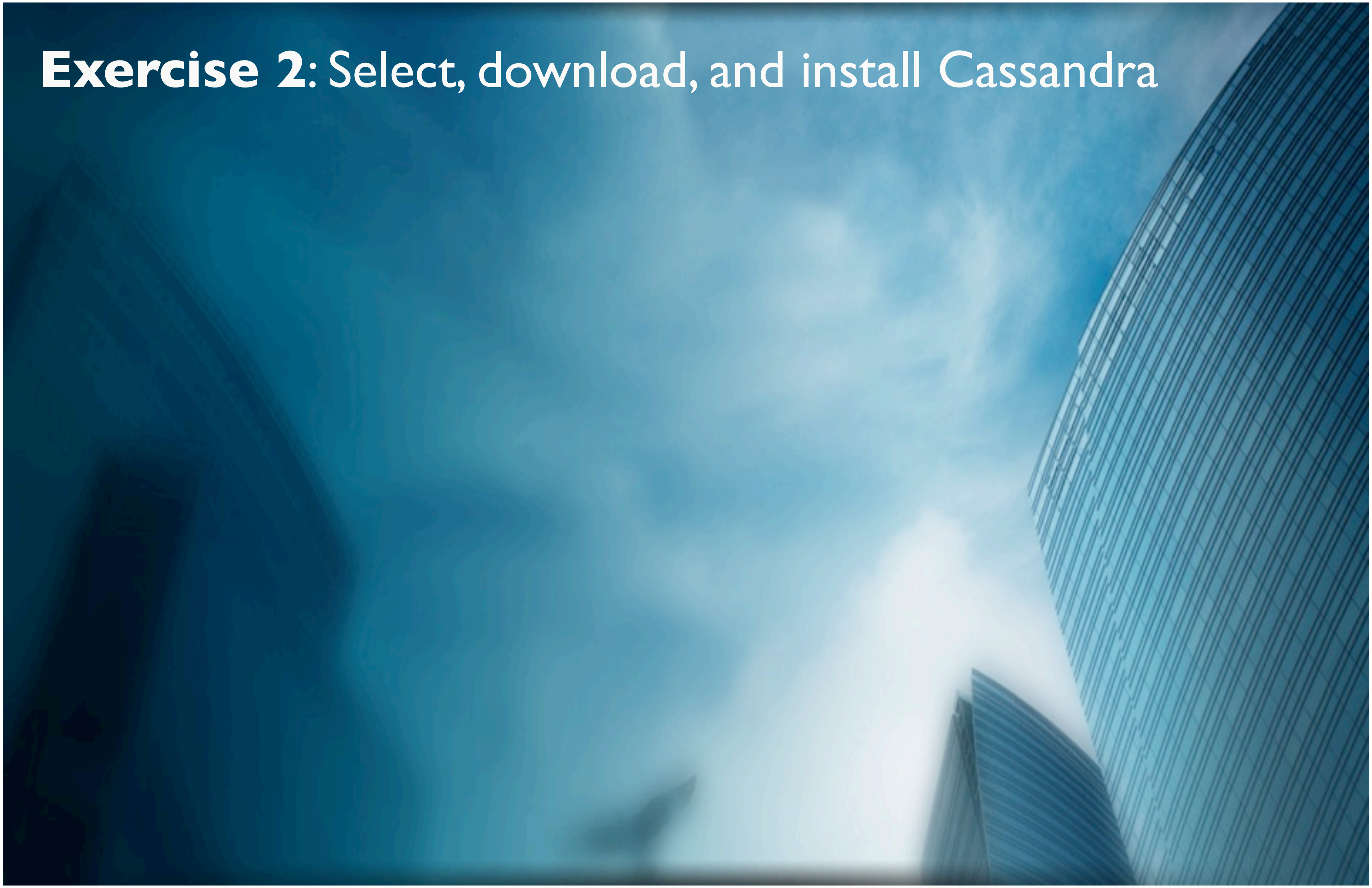
What is in the tarball folder?

- Tarball installations create these folders in a single location:
  - bin: executables (*cassandra*, *cqlsh*, *nodetool*, etc)
  - conf: Configuration files - *cassandra.yaml*
  - javadoc: C* source documentation
  - lib: Library dependencies (jars)
  - pylib: Python libraries (e.g. for *cqlsh*, which is written in Python)
  - tools: Additional tools (e.g. *cassandra-stress* which stresses a C* cluster)



```
Name
▼ apache-cassandra-2.1.0
    ▶ bin
      CHANGES.txt
    ▶ conf
    ▶ interface
    ▶ javadoc
    ▶ lib
      LICENSE.txt
      NEWS.txt
      NOTICE.txt
    ▶ pylib
    ▶ tools
  apache-cassandra-2.1.0-bin.tar.gz
```

Exercise 2: Select, download, and install Cassandra

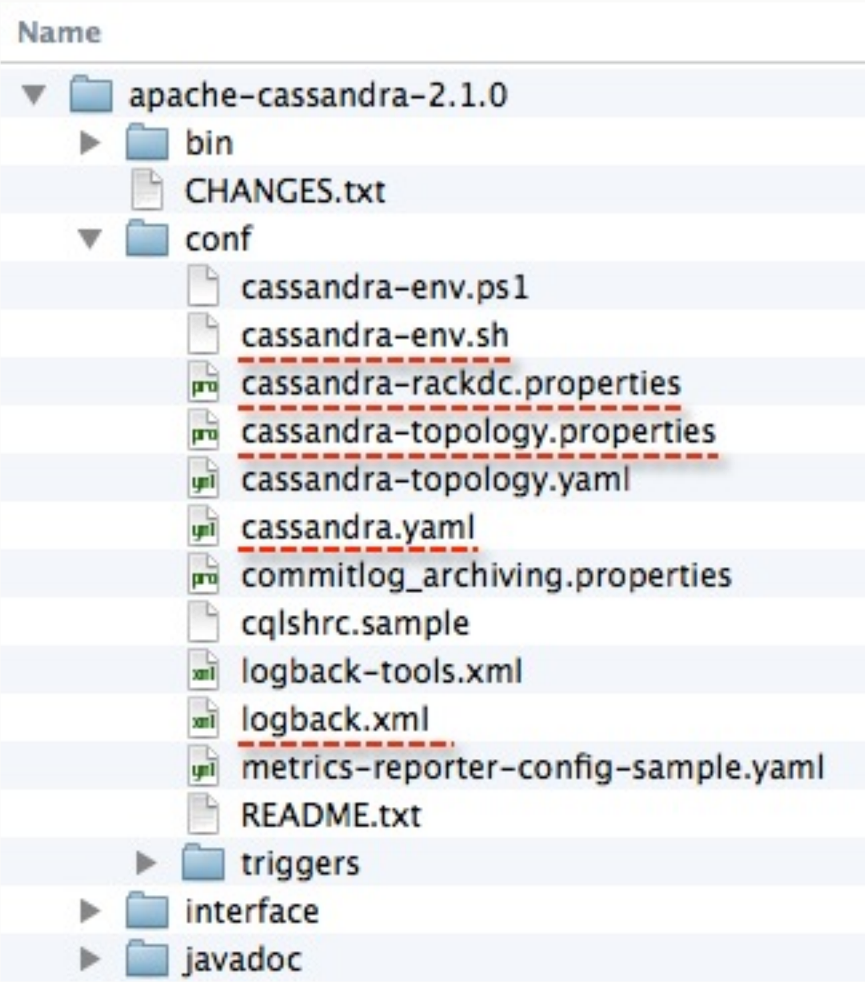


Learning Objectives

- Prepare the operating system
- Select and install a Cassandra distribution
- **Configure Cassandra for a single node**
- Start and stop a Cassandra instance

What configuration files may be used?

- Configuration files include :
  - cassandra.yaml: primary config file for each instance (e.g. data directory locations, etc.)
  - cassandra-env.sh: Java environment config (e.g. MAX_HEAP_SIZE, etc.)
  - logback.xml: system log settings
  - cassandra-rackdc.properties: config to set the Rack and Data Center to which this node belongs.
  - cassandra-topology.properties: config IP addressing for Racks and Data Centers in this cluster
  - bin/cassandra-in.sh: JAVA_HOME, CASSANDRA_CONF, CLASSPATH



```
Name
▼ apache-cassandra-2.1.0
    ▶ bin
      CHANGES.txt
    ▼ conf
        cassandra-env.ps1
        cassandra-env.sh
        cassandra-rackdc.properties
        cassandra-topology.properties
        cassandra-topology.yaml
        cassandra.yaml
        commitlog_archiving.properties
        cqlshrc.sample
        logback-tools.xml
        logback.xml
        metrics-reporter-config-sample.yaml
        README.txt
      ▶ triggers
    ▶ interface
    ▶ javadoc
```

What key properties are set in *cassandra.yaml*?

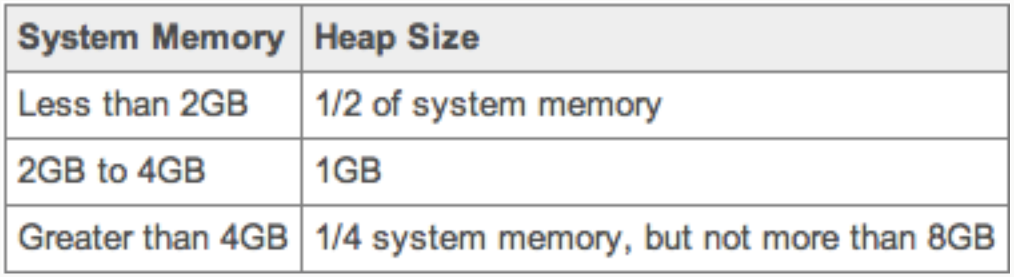
- cluster_name *(default: 'Test Cluster')*
  - All nodes in a cluster must have the same value.
- listen_address *(default: localhost)*
  - IP address or hostname other nodes use to connect to this node
- rpc_address / rpc_port *(default: localhost / 9160)*
  - listen address / port for Thrift client connections
- native_transport_port *(default: 9042)*
  - listen address for Native Java Driver binary protocol

What key properties are set in *cassandra.yaml*?

- commitlog_directory *(default: /var/lib/cassandra/commitlog or $CASSANDRA_HOME/data/commitlog)*
  - Best practice to mount on a separate disk in production (unless SSD)
- data_file_directories *(default: /var/lib/cassandra/data or $CASSANDRA_HOME/data/data)*
  - Storage directory for data tables (SSTables)
- saved_caches_directory *(default: /var/lib/cassandra/saved_caches or $CASSANDRA_HOME/data/saved_caches)*
  - Storage directory for key and row caches

What key properties are set in *cassandra-env.sh*?

- JVM Heap Size settings
  - MAX_HEAP_SIZE="*value*"
    - Maximum recommended in production is currently 8G due to current limitations in Java garbage collection

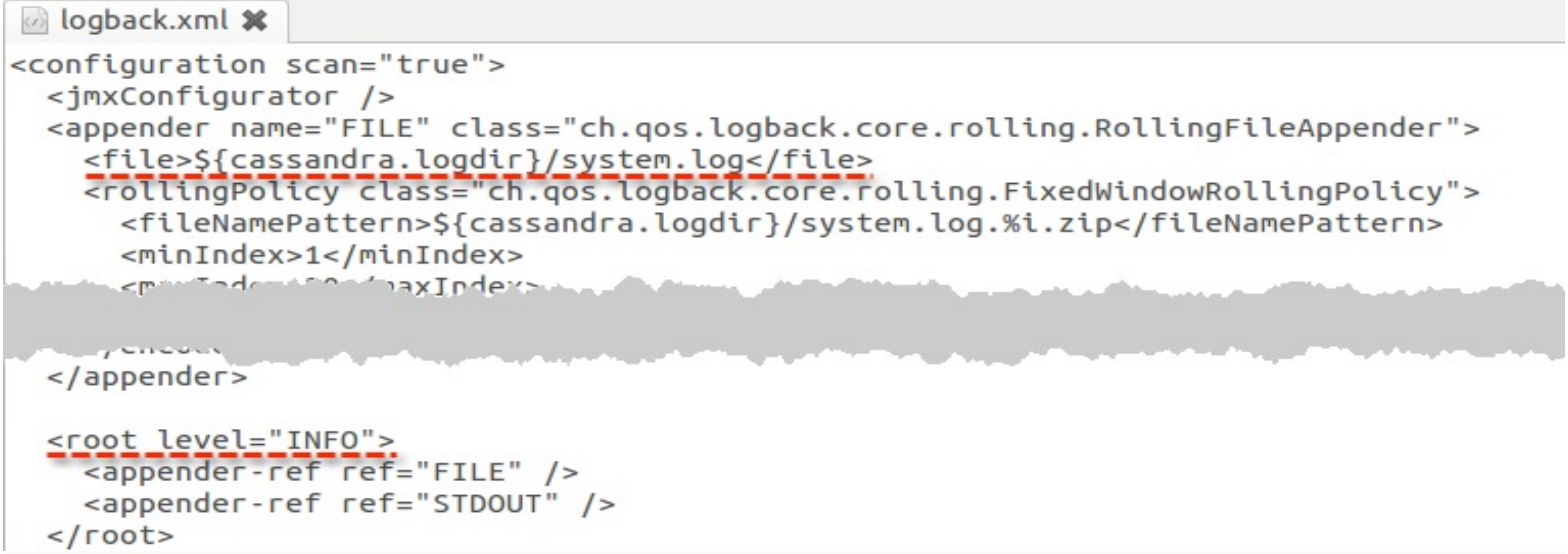


System Memory	Heap Size
Less than 2GB	1/2 of system memory
2GB to 4GB	1GB
Greater than 4GB	1/4 system memory, but not more than 8GB

  - HEAP_NEWSIZE="*value*"
    - Generally set to ¼ of MAX_HEAP_SIZE

What key properties are set in *logback.xml*?

- Cassandra system.log location
  - Default location is `install`*/logs/system.log* (binary tarball) or */var/log/cassandra/system.log* (package install)
  - `system.log` is numerically renamed as it grows over time
- Cassandra logging level
  - Default logging level is INFO



logback.xml

```
<configuration scan="true">
  <jmxConfigurator />
  <appender name="FILE" class="ch.qos.logback.core.rolling.RollingFileAppender">
    <file>${cassandra.logdir}/system.log</file>
    <rollingPolicy class="ch.qos.logback.core.rolling.FixedWindowRollingPolicy">
      <fileNamePattern>${cassandra.logdir}/system.log.%i.zip</fileNamePattern>
      <minIndex>1</minIndex>
      <maxIndex>20</maxIndex>
  </appender>

  <root level="INFO">
    <appender-ref ref="FILE" />
    <appender-ref ref="STDOUT" />
  </root>
```

Exercise 3: Configure a Cassandra instance

DATASTAX

Learning Objectives

- Prepare the operating system
- Select and install a Cassandra distribution
- Configure Cassandra for a single node
- **Start and stop a Cassandra instance**

How do you start Cassandra?

- Launch a server instance using the *cassandra* utility
  - *install*/bin/cassandra

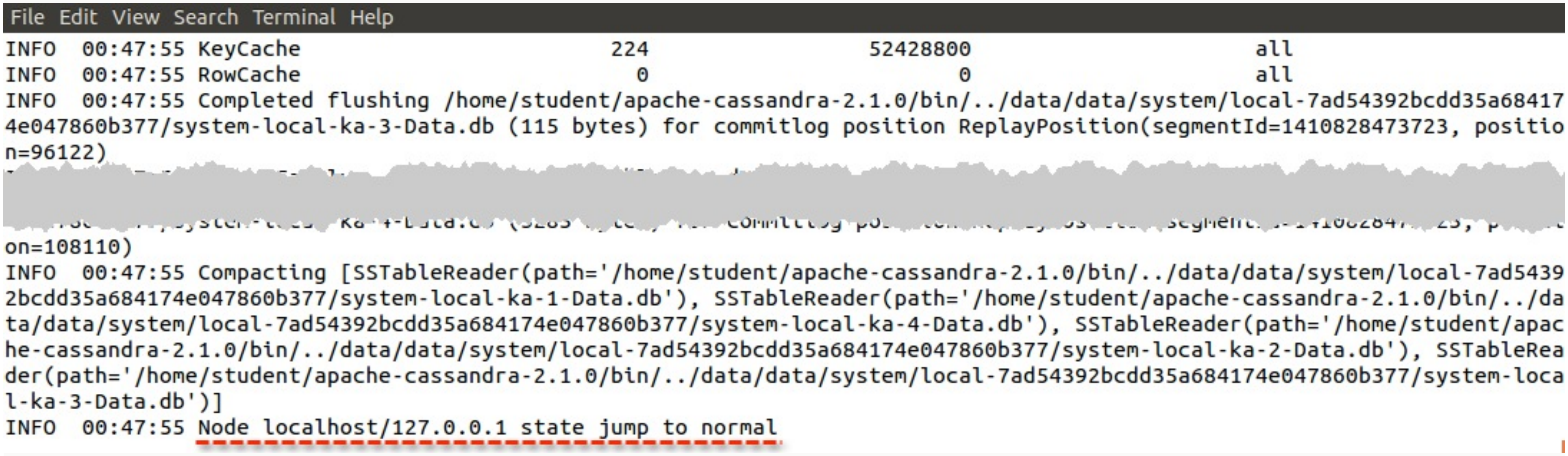
    `cassandra <options>`

    - `-f`
      start Cassandra in foreground (default is background process)
    - `-p <filename>`
      Log process ID in named file; useful to stop Cassandra by PID
    - `-v`
      print the distribution and exit
    - `-D <parameter>`
      Pass a startup parameter (see documentation)

How do you start Cassandra?

- Start a tarball install instance
  - In a terminal window, use *sudo cassandra –f* to launch foreground instance

```
sudo bin/cassandra -f
```



```
INFO  00:47:55 KeyCache                        224                  52428800                       all
INFO  00:47:55 RowCache                          0                         0                       all
INFO  00:47:55 Completed flushing /home/student/apache-cassandra-2.1.0/bin/../data/data/system/local-7ad54392bcdd35a68417
4e047860b377/system-local-ka-3-Data.db (115 bytes) for commitlog position ReplayPosition(segmentId=1410828473723, positio
n=96122)
on=108110)
INFO  00:47:55 Compacting [SSTableReader(path='/home/student/apache-cassandra-2.1.0/bin/../data/data/system/local-7ad5439
2bcdd35a684174e047860b377/system-local-ka-1-Data.db'), SSTableReader(path='/home/student/apache-cassandra-2.1.0/bin/../da
ta/data/system/local-7ad54392bcdd35a684174e047860b377/system-local-ka-4-Data.db'), SSTableReader(path='/home/student/apac
he-cassandra-2.1.0/bin/../data/data/system/local-7ad54392bcdd35a684174e047860b377/system-local-ka-2-Data.db'), SSTableRea
der(path='/home/student/apache-cassandra-2.1.0/bin/../data/data/system/local-7ad54392bcdd35a684174e047860b377/system-loca
l-ka-3-Data.db')]
INFO  00:47:55 Node localhost/127.0.0.1 state jump to normal
```

- Start a package install instance

```
sudo service cassandra start
```

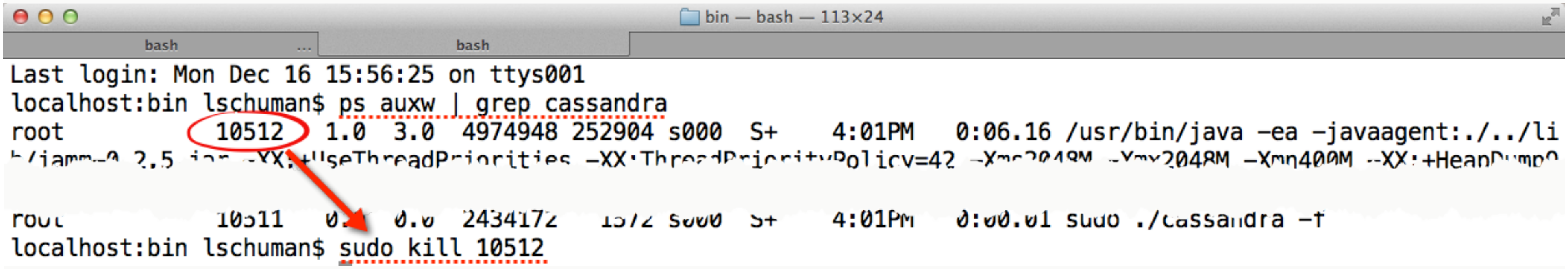
How do you stop Cassandra?

- Stop a tarball install instance – foreground

  `ctrl-c` (in terminal window)

- Stop a tarball install instance – background
  - Determine the Process ID (PID)

  ```
  ps auxw | grep cassandra
  sudo kill <pid>
  ```



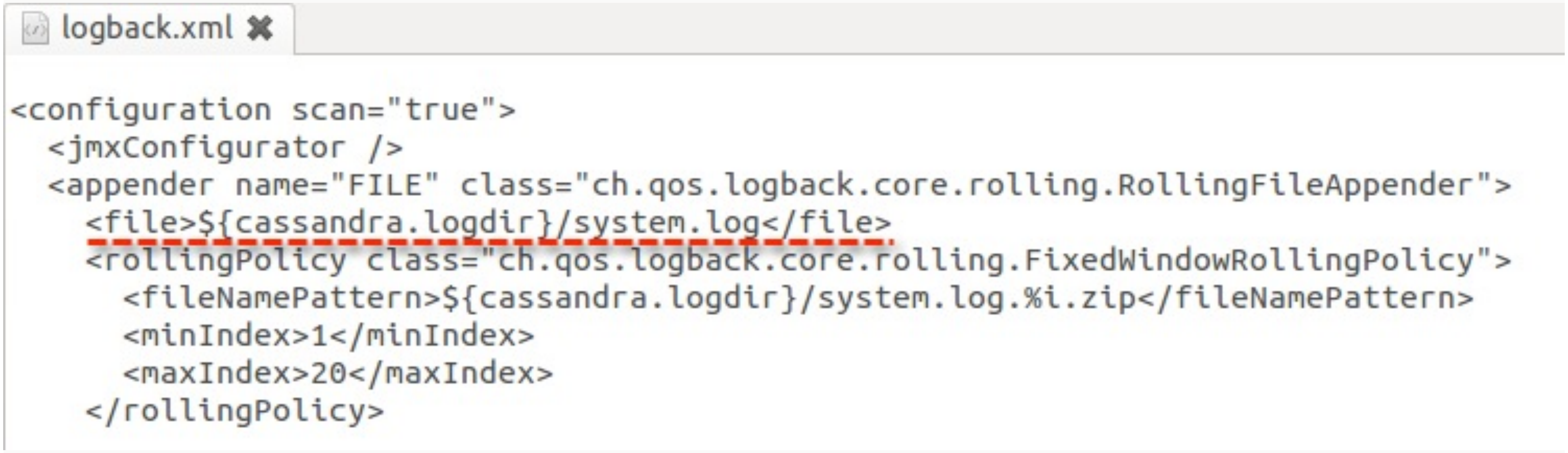
```
Last login: Mon Dec 16 15:56:25 on ttys001
localhost:bin lschuman$ ps auxw | grep cassandra
root            10512  1.0  3.0  4974948 252904 s000  S+    4:01PM   0:06.16 /usr/bin/java -ea -javaagent:./../li
b/jamm-0.2.5.jar -XX:+UseThreadPriorities -XX:ThreadPriorityPolicy=42 -Xms2048M -Xmx2048M -Xmn400M -XX:+HeapDumpO

root            10511  0.0  0.0  2434172   1372 s000  S+    4:01PM   0:00.01 sudo ./cassandra -f
localhost:bin lschuman$ sudo kill 10512
```

- Stop a package install instance

  ```
  sudo service cassandra stop
  ```

How do you locate and review log data?

- System log location set by configuration
  - *install*/conf/logback.xml
- Be sure to distinguish
  - system.log: system state log file, duplicates *stdout*, configurable by logging level
  - CommitLog: table-specific files used during INSERT and UPDATE operations



```
<configuration scan="true">
  <jmxConfigurator />
  <appender name="FILE" class="ch.qos.logback.core.rolling.RollingFileAppender">
    <file>${cassandra.logdir}/system.log</file>
    <rollingPolicy class="ch.qos.logback.core.rolling.FixedWindowRollingPolicy">
      <fileNamePattern>${cassandra.logdir}/system.log.%i.zip</fileNamePattern>
      <minIndex>1</minIndex>
      <maxIndex>20</maxIndex>
    </rollingPolicy>
```

Exercise 4: Run Cassandra and examine its logs

Summary

- *Java 7 JDK* (required) and *JNA* (required) in production
- Synchronize all clocks (NTP) and disable swap
- 3 Cassandra distros: *DS-Enterprise*, *DS-Community*, and *Apache*
- DataStax adds: *OpsCenter*, *DevCenter*, *Security*, *Search*, *Analytics*
- Installation by DEB or YUM package, MSI installer, or tarball
- Configure: *cassandra.yaml*, *cassanda-env.sh, logback.xml, cassandra-rackdc.properties, cassandra-topology.properties*
- Nodes in a cluster share a *cluster_name* and receive connections via their own *listen_address*
- Data is managed in */data*, */commitlog*, and */saved_caches* directories
- Start Cassandra foreground with *bin/cassandra –f*
- Stop Cassandra foreground with *ctrl-c* in its terminal window
- Monitor Cassandra behavior through the *system.log* configured in *logback.xml*

Review Questions

- What open source library must be installed for production use?
- Where do you find and set the system log file location?
- What setting determines a node's cluster, and where is it configured?
- How would you stop a background Cassandra instance on Linux or Mac OSX?
- What settings might you adjust, in which configuration file, to tune Cassandra memory use?

DATASTAX