PD2-Web_Server_Log-MTI850-A25

October 1, 2025

1 PD2: Web Server Log Analysis

1.1 MTI850 - Big Data Analytics

Version 1.0 - Fall 2021

Version 2.0 - Fall 2024

Version 3.0 - Fall 2025 - Apache Spark 4.0

Equipe	9

This assignment will show how easy it is to perform web server log analysis with Apache Spark.

Server log analysis is an ideal use case for Spark. It's a very large, common data source and contains a rich set of information. Spark allows you to store your logs in files on disk cheaply, while still providing a quick and simple way to perform data analysis on them.

This assignment will show you how to use Apache Spark on real-world text-based production logs and fully harness the power of that data.

Log data comes from many sources, such as web, file, and compute servers, application logs, user-generated content, and can be used for monitoring servers, improving business and customer intelligence, building recommendation systems, fraud detection, and much more.

1.1.1 How to complete this assignment

This assignment is broken up into sections examples for demonstrating Spark functionality for log processing.

It consists of 5 parts: * *Part 1*: Introduction and Imports * *Part 2*: Exploratory Data Analysis * *Part 3*: Analysis Walk-Through on the Web Server Log File * *Part 4*: Analyzing Web Server Log File * *Part 5*: Exploring 404 Response Codes

1.2 Part 1: Introduction and Imports

[1]:
```
import findspark
findspark.init()

# Test module for MTI850
```

```
import testmti850
# Util module for MTI850
import utilmti850

import pyspark
from pyspark.sql import SparkSession

spark = SparkSession.builder \
.master("local") \
.appName("Web Server Log Analysis") \
.config("spark.some.config.option", "some-value") \
.getOrCreate()
```

```
WARNING: Using incubator modules: jdk.incubator.vector
Using Spark's default log4j profile: org/apache/spark/log4j2-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use
setLogLevel(newLevel).
25/10/01 12:01:30 WARN NativeCodeLoader: Unable to load native-hadoop library
for your platform… using builtin-java classes where applicable
```

1.2.1 A note about DataFrame column references

In Python, it's possible to access a DataFrame's columns either by attribute (`df.age`) or by indexing (`df['age']`). Referring to a column by attribute (`df.age`) is very Pandas-like, and it's highly convenient, especially when you're doing interactive data exploration. But it can fail, for reasons that aren't obvious. To observe this fact, uncomment the second line and run the next cell (after that, comment again the second line).

[2]:
```
throwaway_df = spark.createDataFrame([('Anthony', 10), ('Julia', 20), ('Fred',␣
↪5)], ('name', 'count'))

throwaway_df.select(throwaway_df.count).show() # This line does not work.␣
↪Please comment it out later.
```

```
---------------------------------------------------------------------------
PySparkTypeError                          Traceback (most recent call last)
Cell In[2], line 3
      1 throwaway_df = spark.createDataFrame([('Anthony', 10), ('Julia', 20),␣
↪('Fred', 5)], ('name', 'count'))
----> 3 throwaway_df.select(throwaway_df.count).show()

File /opt/spark/python/pyspark/sql/classic/dataframe.py:991, in DataFrame.
↪select(self, *cols)
    990 def select(self, *cols: "ColumnOrName") -> ParentDataFrame:  # type:␣
↪ignore[misc]
--> 991     jdf = self._jdf.select(self._jcols(*cols))
    992     return DataFrame(jdf, self.sparkSession)
```

```
File /opt/spark/python/pyspark/sql/classic/dataframe.py:890, in DataFrame.
 ↪_jcols(self, *cols)
    888 if len(cols) == 1 and isinstance(cols[0], list):
    889     cols = cols[0]
--> 890 return self._jseq(cols, _to_java_column)

File /opt/spark/python/pyspark/sql/classic/dataframe.py:877, in DataFrame.
 ↪_jseq(self, cols, converter)
    871 def _jseq(
    872     self,
    873     cols: Sequence,
    874     converter: Optional[Callable[..., Union["PrimitiveType",␣
 ↪"JavaObject"]]] = None,
    875 ) -> "JavaObject":
    876     """Return a JVM Seq of Columns from a list of Column or names"""
--> 877     return _to_seq(self.sparkSession._sc, cols, converter)

File /opt/spark/python/pyspark/sql/classic/column.py:104, in _to_seq(sc, cols,␣
 ↪converter)
     97 """
     98 Convert a list of Columns (or names) into a JVM Seq of Column.
     99
    100 An optional `converter` could be used to convert items in `cols`
    101 into JVM Column objects.
    102 """
    103 if converter:
--> 104     cols = [converter(c) for c in cols]
    105 assert sc._jvm is not None
    106 return sc._jvm.PythonUtils.toSeq(cols)

File /opt/spark/python/pyspark/sql/classic/column.py:71, in _to_java_column(col
     69     jcol = _create_column_from_name(col)
     70 else:
---> 71     raise PySparkTypeError(
     72         errorClass="NOT_COLUMN_OR_STR",
     73         messageParameters={"arg_name": "col", "arg_type": type(col).
 ↪__name__},
     74     )
     75 return jcol

PySparkTypeError: [NOT_COLUMN_OR_STR] Argument `col` should be a Column or str,
 ↪got method.
```

To understand why that failed, you have to understand how the attribute-column syntax is implemented.

When you type `throwaway_df.count`, Python looks for an *existing* attribute or method called

`count` on the `throwaway_df` object. If it finds one, it uses it. Otherwise, it calls a special Python function (`__getattr__`), which defaults to throwing an exception. Spark has overridden `__getattr__` to look for a column on the DataFrame.

This means you can only use the attribute (dot) syntax to refer to a column if the DataFrame does not *already* have an attribute with the column's name.

In the above example, there's already a `count()` method on the `DataFrame` class, so `throwaway_df.count` does not refer to our "count" column; instead, it refers to the `count()` *method*.

To avoid this problem, you can refer to the column using subscript notation: `throwaway_df['count']`. This syntax will *always* work.

[7]:
```
throwaway_df.select(throwaway_df['count']).show()
```

```
+-----+
|count|
+-----+
|   10|
|   20|
|    5|
+-----+
```

1.2.2 (1a) Library Imports

We can import standard Python3 libraries (modules) the usual way. An `import` statement will import the specified module. In this assignment, we will provide any imports that are necessary.

Let's import some of the libraries we'll need:

- `re`: The regular expression library
- `datetime`: Date and time functions

[8]:
```
import re
import datetime
```

[9]:
```
# Quick test of the regular expression library
m = re.search('(?<=abc)def', 'abcdef')
m.group(0)
```

[9]:
```
'def'
```

[10]:
```
# Quick test of the datetime library
print ('This was last run on: {0}'.format(datetime.datetime.now()))
```

```
This was last run on: 2025-10-01 08:06:48.069034
```

1.2.3 (1b) Getting help

Remember: There are some useful Python built-ins for getting help.

You can use Python's dir() function to get a list of all the attributes (including methods) accessible through the `spark` object.

[11]:
```
dir(spark)
```

[11]:
```
['Builder',
 '__annotations__',
 '__class__',
 '__delattr__',
 '__dict__',
 '__dir__',
 '__doc__',
 '__enter__',
 '__eq__',
 '__exit__',
 '__firstlineno__',
 '__format__',
 '__ge__',
 '__getattribute__',
 '__getstate__',
 '__gt__',
 '__hash__',
 '__init__',
 '__init_subclass__',
 '__le__',
 '__lt__',
 '__module__',
 '__ne__',
 '__new__',
 '__reduce__',
 '__reduce_ex__',
 '__repr__',
 '__setattr__',
 '__sizeof__',
 '__static_attributes__',
 '__str__',
 '__subclasshook__',
 '__weakref__',
 '_activeSession',
 '_convert_from_pandas',
 '_createFromLocal',
 '_createFromRDD',
 '_create_dataframe',
 '_create_from_arrow_table',
 '_create_from_pandas_with_arrow',
 '_create_shell_session',
 '_getActiveSessionOrCreate',
```

```
'_get_j_spark_session_class',
'_get_j_spark_session_module',
'_get_numpy_record_dtype',
'_inferSchema',
'_inferSchemaFromList',
'_instantiatedSession',
'_jconf',
'_jsc',
'_jsparkSession',
'_jvm',
'_parse_ddl',
'_profiler_collector',
'_repr_html_',
'_sc',
'_should_update_active_session',
'_to_ddl',
'active',
'addArtifact',
'addArtifacts',
'addTag',
'builder',
'catalog',
'clearProgressHandlers',
'clearTags',
'client',
'conf',
'copyFromLocalToFs',
'createDataFrame',
'dataSource',
'getActiveSession',
'getTags',
'interruptAll',
'interruptOperation',
'interruptTag',
'newSession',
'profile',
'range',
'read',
'readStream',
'registerProgressHandler',
'removeProgressHandler',
'removeTag',
'sparkContext',
'sql',
'stop',
'streams',
'table',
```

```
 'tvf',
 'udf',
 'udtf',
 'version']
```

Alternatively, you can use Python's help() function to get an easier to read list of all the attributes, including examples, that the `spark` object has.

[12]:
```
# Use help to obtain more detailed information
help(spark)
```

```
Help on SparkSession in module pyspark.sql.session object:

class SparkSession(pyspark.sql.pandas.conversion.SparkConversionMixin)
 |  SparkSession(
 |      sparkContext: 'SparkContext',
 |      jsparkSession: Optional[ForwardRef('JavaObject')] = None,
 |      options: Dict[str, Any] = {}
 |  )
 |
 |  The entry point to programming Spark with the Dataset and DataFrame API.
 |
 |  A SparkSession can be used to create :class:`DataFrame`, register
:class:`DataFrame` as
 |  tables, execute SQL over tables, cache tables, and read parquet files.
 |  To create a :class:`SparkSession`, use the following builder pattern:
 |
 |  .. versionchanged:: 3.4.0
 |      Supports Spark Connect.
 |
 |  .. autoattribute:: builder
 |     :annotation:
 |
 |  Examples
 |  --------
 |  Create a Spark session.
 |
 |  >>> spark = (
 |  …     SparkSession.builder
 |  …         .master("local")
 |  …         .appName("Word Count")
 |  …         .config("spark.some.config.option", "some-value")
 |  …         .getOrCreate()
 |  … )
 |
 |  Create a Spark session with Spark Connect.
 |
 |  >>> spark = (
```

```
 |  …     SparkSession.builder
 |  …         .remote("sc://localhost")
 |  …         .appName("Word Count")
 |  …         .config("spark.some.config.option", "some-value")
 |  …         .getOrCreate()
 |  … )  # doctest: +SKIP
 |
 |  Method resolution order:
 |      SparkSession
 |      pyspark.sql.pandas.conversion.SparkConversionMixin
 |      builtins.object
 |
 |  Methods defined here:
 |
 |  __enter__(self) -> 'SparkSession'
 |      Enable 'with SparkSession.builder.(…).getOrCreate() as session: app'
syntax.
 |
 |      .. versionadded:: 2.0.0
 |
 |      Examples
 |      --------
 |      >>> with SparkSession.builder.master("local").getOrCreate() as session:
 |      …     session.range(5).show()  # doctest: +SKIP
 |      +---+
 |      | id|
 |      +---+
 |      |  0|
 |      |  1|
 |      |  2|
 |      |  3|
 |      |  4|
 |      +---+
 |
 |  __exit__(
 |      self,
 |      exc_type: Optional[Type[BaseException]],
 |      exc_val: Optional[BaseException],
 |      exc_tb: Optional[traceback]
 |  ) -> None
 |      Enable 'with SparkSession.builder.(…).getOrCreate() as session: app'
syntax.
 |
 |      Specifically stop the SparkSession on exit of the with block.
 |
 |      .. versionadded:: 2.0.0
 |
 |      Examples
```

```
 |      --------
 |      >>> with SparkSession.builder.master("local").getOrCreate() as session:
 |      …     session.range(5).show()  # doctest: +SKIP
 |      +---+
 |      | id|
 |      +---+
 |      |  0|
 |      |  1|
 |      |  2|
 |      |  3|
 |      |  4|
 |      +---+
 |
 |  __init__(
 |      self,
 |      sparkContext: 'SparkContext',
 |      jsparkSession: Optional[ForwardRef('JavaObject')] = None,
 |      options: Dict[str, Any] = {}
 |  )
 |      Initialize self.  See help(type(self)) for accurate signature.
 |
 |  addArtifact = addArtifacts(
 |      self,
 |      *path: str,
 |      pyfile: bool = False,
 |      archive: bool = False,
 |      file: bool = False
 |  ) -> None
 |
 |  addArtifacts(
 |      self,
 |      *path: str,
 |      pyfile: bool = False,
 |      archive: bool = False,
 |      file: bool = False
 |  ) -> None
 |      Add artifact(s) to the client session. Currently only local files are
supported.
 |
 |      .. versionadded:: 3.5.0
 |
 |      .. versionchanged:: 4.0.0
 |          Supports Spark Classic.
 |
 |      Parameters
 |      ----------
 |      *path : tuple of str
 |          Artifact's URIs to add.
```

```
 |      pyfile : bool
 |          Whether to add them as Python dependencies such as .py, .egg, .zip
or .jar files.
 |          The pyfiles are directly inserted into the path when executing
Python functions
 |          in executors.
 |      archive : bool
 |          Whether to add them as archives such as .zip, .jar, .tar.gz, .tgz,
or .tar files.
 |          The archives are unpacked on the executor side automatically.
 |      file : bool
 |          Add a file to be downloaded with this Spark job on every node.
 |          The ``path`` passed can only be a local file for now.
 |
 |  addTag(self, tag: str) -> None
 |      Add a tag to be assigned to all the operations started by this thread in
this session.
 |
 |      Often, a unit of execution in an application consists of multiple Spark
executions.
 |      Application programmers can use this method to group all those jobs
together and give a
 |      group tag. The application can use :meth:`SparkSession.interruptTag` to
cancel all running
 |      executions with this tag.
 |
 |      There may be multiple tags present at the same time, so different parts
of application may
 |      use different tags to perform cancellation at different levels of
granularity.
 |
 |      .. versionadded:: 3.5.0
 |
 |      .. versionchanged:: 4.0.0
 |          Supports Spark Classic.
 |
 |      Parameters
 |      ----------
 |      tag : str
 |          The tag to be added. Cannot contain ',' (comma) character or be an
empty string.
 |
 |  catalog = <functools.cached_property object>
 |      Interface through which the user may create, drop, alter or query
underlying
 |      databases, tables, functions, etc.
 |
 |      .. versionadded:: 2.0.0
```

```
 |
 |      .. versionchanged:: 3.4.0
 |          Supports Spark Connect.
 |
 |      Returns
 |      -------
 |      :class:`Catalog`
 |
 |      Examples
 |      --------
 |      >>> spark.catalog
 |      <…Catalog object …>
 |
 |      Create a temp view, show the list, and drop it.
 |
 |      >>> spark.range(1).createTempView("test_view")
 |      >>> spark.catalog.listTables()  # doctest: +SKIP
 |      [Table(name='test_view', catalog=None, namespace=[], description=None,
…
 |      >>> _ = spark.catalog.dropTempView("test_view")
 |
 |  clearProgressHandlers(self) -> None
 |      Clear all registered progress handlers.
 |
 |      .. versionadded:: 4.0
 |
 |  clearTags(self) -> None
 |      Clear the current thread's operation tags.
 |
 |      .. versionadded:: 3.5.0
 |
 |      .. versionchanged:: 4.0.0
 |          Supports Spark Classic.
 |
 |  conf = <functools.cached_property object>
 |      Runtime configuration interface for Spark.
 |
 |      This is the interface through which the user can get and set all Spark
and Hadoop
 |      configurations that are relevant to Spark SQL. When getting the value of
a config,
 |      this defaults to the value set in the underlying :class:`SparkContext`,
if any.
 |
 |      .. versionadded:: 2.0.0
 |
 |      .. versionchanged:: 3.4.0
 |          Supports Spark Connect.
```

```
 |
 |      Returns
 |      -------
 |      :class:`pyspark.sql.conf.RuntimeConfig`
 |
 |      Examples
 |      --------
 |      >>> spark.conf
 |      <pyspark…RuntimeConf…>
 |
 |      Set a runtime configuration for the session
 |
 |      >>> spark.conf.set("key", "value")
 |      >>> spark.conf.get("key")
 |      'value'
 |
 |  copyFromLocalToFs(self, local_path: str, dest_path: str) -> None
 |      Copy file from local to cloud storage file system.
 |      If the file already exits in destination path, old file is overwritten.
 |
 |      .. versionadded:: 3.5.0
 |
 |      Parameters
 |      ----------
 |      local_path: str
 |          Path to a local file. Directories are not supported.
 |          The path can be either an absolute path or a relative path.
 |      dest_path: str
 |          The cloud storage path to the destination the file will
 |          be copied to.
 |          The path must be an an absolute path.
 |
 |      Notes
 |      -----
 |      This API is a developer API.
 |      Also, this is an API dedicated to Spark Connect client only. With
regular
 |      Spark Session, it throws an exception.
 |
 |  createDataFrame(
 |      self,
 |      data: Union[ForwardRef('RDD[Any]'), Iterable[Any],
ForwardRef('PandasDataFrameLike'), ForwardRef('ArrayLike'),
ForwardRef('pa.Table')],
 |      schema: Union[pyspark.sql.types.AtomicType,
pyspark.sql.types.StructType, str, NoneType] = None,
 |      samplingRatio: Optional[float] = None,
 |      verifySchema: bool = True
```

```
 |  ) -> pyspark.sql.dataframe.DataFrame
 |      Creates a :class:`DataFrame` from an :class:`RDD`, a list, a
:class:`pandas.DataFrame`,
 |      a :class:`numpy.ndarray`, or a :class:`pyarrow.Table`.
 |
 |      .. versionadded:: 2.0.0
 |
 |      .. versionchanged:: 3.4.0
 |          Supports Spark Connect.
 |
 |      .. versionchanged:: 4.0.0
 |          Supports :class:`pyarrow.Table`.
 |
 |      Parameters
 |      ----------
 |      data : :class:`RDD` or iterable
 |          an RDD of any kind of SQL data representation (:class:`Row`,
 |          :class:`tuple`, ``int``, ``boolean``, ``dict``, etc.), or
:class:`list`,
 |          :class:`pandas.DataFrame`, :class:`numpy.ndarray`, or
:class:`pyarrow.Table`.
 |      schema : :class:`pyspark.sql.types.DataType`, str or list, optional
 |          a :class:`pyspark.sql.types.DataType` or a datatype string or a list
of
 |          column names, default is None. The data type string format equals to
 |          :class:`pyspark.sql.types.DataType.simpleString`, except that top
level struct type can
 |          omit the ``struct<>``.
 |
 |          When ``schema`` is a list of column names, the type of each column
 |          will be inferred from ``data``.
 |
 |          When ``schema`` is ``None``, it will try to infer the schema (column
names and types)
 |          from ``data``, which should be an RDD of either :class:`Row`,
 |          :class:`namedtuple`, or :class:`dict`.
 |
 |          When ``schema`` is :class:`pyspark.sql.types.DataType` or a datatype
string, it must
 |          match the real data, or an exception will be thrown at runtime. If
the given schema is
 |          not :class:`pyspark.sql.types.StructType`, it will be wrapped into a
 |          :class:`pyspark.sql.types.StructType` as its only field, and the
field name will be
 |          "value". Each record will also be wrapped into a tuple, which can be
converted to row
 |          later.
 |      samplingRatio : float, optional
```

```
 |          the sample ratio of rows used for inferring. The first few rows will
be used
 |          if ``samplingRatio`` is ``None``. This option is effective only when
the input is
 |          :class:`RDD`.
 |      verifySchema : bool, optional
 |          verify data types of every row against schema. Enabled by default.
 |          When the input is :class:`pyarrow.Table` or when the input class is
 |          :class:`pandas.DataFrame` and
`spark.sql.execution.arrow.pyspark.enabled` is enabled,
 |          this option is not effective. It follows Arrow type coercion. This
option is not
 |          supported with Spark Connect.
 |
 |          .. versionadded:: 2.1.0
 |
 |      Returns
 |      -------
 |      :class:`DataFrame`
 |
 |      Notes
 |      -----
 |      Usage with `spark.sql.execution.arrow.pyspark.enabled=True` is
experimental.
 |
 |      Examples
 |      --------
 |      Create a DataFrame from a list of tuples.
 |
 |      >>> spark.createDataFrame([('Alice', 1)]).show()
 |      +-----+---+
 |      |   _1| _2|
 |      +-----+---+
 |      |Alice|  1|
 |      +-----+---+
 |
 |      Create a DataFrame from a list of dictionaries.
 |
 |      >>> d = [{'name': 'Alice', 'age': 1}]
 |      >>> spark.createDataFrame(d).show()
 |      +---+-----+
 |      |age| name|
 |      +---+-----+
 |      |  1|Alice|
 |      +---+-----+
 |
 |      Create a DataFrame with column names specified.
 |
```

```
 |      >>> spark.createDataFrame([('Alice', 1)], ['name', 'age']).show()
 |      +-----+---+
 |      | name|age|
 |      +-----+---+
 |      |Alice|  1|
 |      +-----+---+
 |
 |      Create a DataFrame with the explicit schema specified.
 |
 |      >>> from pyspark.sql.types import *
 |      >>> schema = StructType([
 |      …    StructField("name", StringType(), True),
 |      …    StructField("age", IntegerType(), True)])
 |      >>> spark.createDataFrame([('Alice', 1)], schema).show()
 |      +-----+---+
 |      | name|age|
 |      +-----+---+
 |      |Alice|  1|
 |      +-----+---+
 |
 |      Create a DataFrame with the schema in DDL formatted string.
 |
 |      >>> spark.createDataFrame([('Alice', 1)], "name: string, age:
int").show()
 |      +-----+---+
 |      | name|age|
 |      +-----+---+
 |      |Alice|  1|
 |      +-----+---+
 |
 |      Create an empty DataFrame.
 |      When initializing an empty DataFrame in PySpark, it's mandatory to
specify its schema,
 |      as the DataFrame lacks data from which the schema can be inferred.
 |
 |      >>> spark.createDataFrame([], "name: string, age: int").show()
 |      +----+---+
 |      |name|age|
 |      +----+---+
 |      +----+---+
 |
 |      Create a DataFrame from Row objects.
 |
 |      >>> from pyspark.sql import Row
 |      >>> Person = Row('name', 'age')
 |      >>> df = spark.createDataFrame([Person("Alice", 1)])
 |      >>> df.show()
 |      +-----+---+
```

```
 |      | name|age|
 |      +-----+---+
 |      |Alice|  1|
 |      +-----+---+
 |
 |      Create a DataFrame from a pandas DataFrame.
 |
 |      >>> spark.createDataFrame(df.toPandas()).show()  # doctest: +SKIP
 |      +-----+---+
 |      | name|age|
 |      +-----+---+
 |      |Alice|  1|
 |      +-----+---+
 |
 |      >>> pdf = pandas.DataFrame([[1, 2]])  # doctest: +SKIP
 |      >>> spark.createDataFrame(pdf).show()  # doctest: +SKIP
 |      +---+---+
 |      |  0|  1|
 |      +---+---+
 |      |  1|  2|
 |      +---+---+
 |
 |      Create a DataFrame from a PyArrow Table.
 |
 |      >>> spark.createDataFrame(df.toArrow()).show()  # doctest: +SKIP
 |      +-----+---+
 |      | name|age|
 |      +-----+---+
 |      |Alice|  1|
 |      +-----+---+
 |
 |      >>> table = pyarrow.table({'0': [1], '1': [2]})  # doctest: +SKIP
 |      >>> spark.createDataFrame(table).show()  # doctest: +SKIP
 |      +---+---+
 |      |  0|  1|
 |      +---+---+
 |      |  1|  2|
 |      +---+---+
 |
 |  getTags(self) -> Set[str]
 |      Get the tags that are currently set to be assigned to all the operations
started by this
 |      thread.
 |
 |      .. versionadded:: 3.5.0
 |
 |      .. versionchanged:: 4.0.0
 |          Supports Spark Classic.
```

```
 |
 |      Returns
 |      -------
 |      set of str
 |          Set of tags of interrupted operations.
 |
 |  interruptAll(self) -> List[str]
 |      Interrupt all operations of this session currently running on the
connected server.
 |
 |      .. versionadded:: 3.5.0
 |
 |      .. versionchanged:: 4.0.0
 |          Supports Spark Classic.
 |
 |      Returns
 |      -------
 |      list of str
 |          List of operationIds of interrupted operations.
 |
 |      Notes
 |      -----
 |      There is still a possibility of operation finishing just as it is
interrupted.
 |
 |  interruptOperation(self, op_id: str) -> List[str]
 |      Interrupt an operation of this session with the given operationId.
 |
 |      .. versionadded:: 3.5.0
 |
 |      .. versionchanged:: 4.0.0
 |          Supports Spark Classic.
 |
 |      Returns
 |      -------
 |      list of str
 |          List of operationIds of interrupted operations.
 |
 |      Notes
 |      -----
 |      There is still a possibility of operation finishing just as it is
interrupted.
 |
 |  interruptTag(self, tag: str) -> List[str]
 |      Interrupt all operations of this session with the given operation tag.
 |
 |      .. versionadded:: 3.5.0
 |
```

```
 |      .. versionchanged:: 4.0.0
 |          Supports Spark Classic.
 |
 |      Returns
 |      -------
 |      list of str
 |          List of operationIds of interrupted operations.
 |
 |      Notes
 |      -----
 |      There is still a possibility of operation finishing just as it is
interrupted.
 |
 |  newSession(self) -> 'SparkSession'
 |      Returns a new :class:`SparkSession` as new session, that has separate
SQLConf,
 |      registered temporary views and UDFs, but shared :class:`SparkContext`
and
 |      table cache.
 |
 |      .. versionadded:: 2.0.0
 |
 |      Returns
 |      -------
 |      :class:`SparkSession`
 |          Spark session if an active session exists for the current thread
 |
 |      Examples
 |      --------
 |      >>> spark.newSession()
 |      <…SparkSession object …>
 |
 |  range(
 |      self,
 |      start: int,
 |      end: Optional[int] = None,
 |      step: int = 1,
 |      numPartitions: Optional[int] = None
 |  ) -> pyspark.sql.dataframe.DataFrame
 |      Create a :class:`DataFrame` with single
:class:`pyspark.sql.types.LongType` column named
 |      ``id``, containing elements in a range from ``start`` to ``end``
(exclusive) with
 |      step value ``step``.
 |
 |      .. versionadded:: 2.0.0
 |
 |      .. versionchanged:: 3.4.0
```

```
 |          Supports Spark Connect.
 |
 |      Parameters
 |      ----------
 |      start : int
 |          the start value
 |      end : int, optional
 |          the end value (exclusive)
 |      step : int, optional
 |          the incremental step (default: 1)
 |      numPartitions : int, optional
 |          the number of partitions of the DataFrame
 |
 |      Returns
 |      -------
 |      :class:`DataFrame`
 |
 |      Examples
 |      --------
 |      >>> spark.range(1, 7, 2).show()
 |      +---+
 |      | id|
 |      +---+
 |      |  1|
 |      |  3|
 |      |  5|
 |      +---+
 |
 |      If only one argument is specified, it will be used as the end value.
 |
 |      >>> spark.range(3).show()
 |      +---+
 |      | id|
 |      +---+
 |      |  0|
 |      |  1|
 |      |  2|
 |      +---+
 |
 |  registerProgressHandler(self, handler: 'ProgressHandler') -> None
 |      Register a progress handler to be called when a progress update is
received from the server.
 |
 |      .. versionadded:: 4.0
 |
 |      Parameters
 |      ----------
 |      handler : ProgressHandler
```

```
 |        A callable that follows the ProgressHandler interface. This handler
will be called
 |        on every progress update.
 |
 |      Examples
 |      --------
 |
 |      >>> def progress_handler(stages, inflight_tasks, done):
 |      …     print(f"{len(stages)} Stages known, Done: {done}")
 |      >>> spark.registerProgressHandler(progress_handler)
 |      >>> res = spark.range(10).repartition(1).collect()  # doctest: +SKIP
 |      3 Stages known, Done: False
 |      3 Stages known, Done: True
 |      >>> spark.clearProgressHandlers()
 |
 |  removeProgressHandler(self, handler: 'ProgressHandler') -> None
 |      Remove a progress handler that was previously registered.
 |
 |      .. versionadded:: 4.0
 |
 |      Parameters
 |      ----------
 |      handler : ProgressHandler
 |        The handler to remove if present in the list of progress handlers.
 |
 |  removeTag(self, tag: str) -> None
 |      Remove a tag previously added to be assigned to all the operations
started by this thread in
 |      this session. Noop if such a tag was not added earlier.
 |
 |      .. versionadded:: 3.5.0
 |
 |      .. versionchanged:: 4.0.0
 |          Supports Spark Classic.
 |
 |      Parameters
 |      ----------
 |      tag : list of str
 |          The tag to be removed. Cannot contain ',' (comma) character or be an
empty string.
 |
 |  sql(
 |      self,
 |      sqlQuery: str,
 |      args: Union[Dict[str, Any], List, NoneType] = None,
 |      **kwargs: Any
 |  ) -> 'ParentDataFrame'
 |      Returns a :class:`DataFrame` representing the result of the given query.
```

```
 |      When ``kwargs`` is specified, this method formats the given string by
using the Python
 |      standard formatter. The method binds named parameters to SQL literals or
 |      positional parameters from `args`. It doesn't support named and
positional parameters
 |      in the same SQL query.
 |
 |      .. versionadded:: 2.0.0
 |
 |      .. versionchanged:: 3.4.0
 |          Supports Spark Connect and parameterized SQL.
 |
 |      .. versionchanged:: 3.5.0
 |          Added positional parameters.
 |
 |      Parameters
 |      ----------
 |      sqlQuery : str
 |          SQL query string.
 |      args : dict or list
 |          A dictionary of parameter names to Python objects or a list of
Python objects
 |          that can be converted to SQL literal expressions. See
 |          `Supported Data Types`_ for supported value types in Python.
 |          For example, dictionary keys: "rank", "name", "birthdate";
 |          dictionary or list values: 1, "Steven", datetime.date(2023, 4, 2).
 |          A value can be also a `Column` of a literal or collection
constructor functions such
 |          as `map()`, `array()`, `struct()`, in that case it is taken as is.
 |
 |          .. _Supported Data Types: https://spark.apache.org/docs/latest/sql-
ref-datatypes.html
 |
 |          .. versionadded:: 3.4.0
 |
 |      kwargs : dict
 |          Other variables that the user wants to set that can be referenced in
the query
 |
 |          .. versionchanged:: 3.3.0
 |             Added optional argument ``kwargs`` to specify the mapping of
variables in the query.
 |             This feature is experimental and unstable.
 |
 |      Returns
 |      -------
 |      :class:`DataFrame`
 |
```

```
 |      Notes
 |      -----
 |      In Spark Classic, a temporary view referenced in `spark.sql` is resolved
immediately,
 |      while in Spark Connect it is lazily analyzed.
 |      So in Spark Connect if a view is dropped, modified or replaced after
`spark.sql`, the
 |      execution may fail or generate different results.
 |
 |      Examples
 |      --------
 |      Executing a SQL query.
 |
 |      >>> spark.sql("SELECT * FROM range(10) where id > 7").show()
 |      +---+
 |      | id|
 |      +---+
 |      |  8|
 |      |  9|
 |      +---+
 |
 |      Executing a SQL query with variables as Python formatter standard.
 |
 |      >>> spark.sql(
 |      …     "SELECT * FROM range(10) WHERE id > {bound1} AND id < {bound2}",
bound1=7, bound2=9
 |      … ).show()
 |      +---+
 |      | id|
 |      +---+
 |      |  8|
 |      +---+
 |
 |      >>> mydf = spark.range(10)
 |      >>> spark.sql(
 |      …     "SELECT {col} FROM {mydf} WHERE id IN {x}",
 |      …     col=mydf.id, mydf=mydf, x=tuple(range(4))).show()
 |      +---+
 |      | id|
 |      +---+
 |      |  0|
 |      |  1|
 |      |  2|
 |      |  3|
 |      +---+
 |
 |      >>> spark.sql('''
 |      …   SELECT m1.a, m2.b
```

```
 |      …   FROM {table1} m1 INNER JOIN {table2} m2
 |      …   ON m1.key = m2.key
 |      …   ORDER BY m1.a, m2.b''',
 |      …   table1=spark.createDataFrame([(1, "a"), (2, "b")], ["a", "key"]),
 |      …   table2=spark.createDataFrame([(3, "a"), (4, "b"), (5, "b")], ["b",
"key"])).show()
 |      +---+---+
 |      |  a|  b|
 |      +---+---+
 |      |  1|  3|
 |      |  2|  4|
 |      |  2|  5|
 |      +---+---+
 |
 |      Also, it is possible to query using class:`Column` from
:class:`DataFrame`.
 |
 |      >>> mydf = spark.createDataFrame([(1, 4), (2, 4), (3, 6)], ["A", "B"])
 |      >>> spark.sql("SELECT {df.A}, {df[B]} FROM {df}", df=mydf).show()
 |      +---+---+
 |      |  A|  B|
 |      +---+---+
 |      |  1|  4|
 |      |  2|  4|
 |      |  3|  6|
 |      +---+---+
 |
 |      And substitute named parameters with the `:` prefix by SQL literals.
 |
 |      >>> from pyspark.sql.functions import create_map, lit
 |      >>> spark.sql(
 |      …   "SELECT *, element_at(:m, 'a') AS C FROM {df} WHERE {df[B]} >
:minB",
 |      …   {"minB" : 5, "m" : create_map(lit('a'), lit(1))}, df=mydf).show()
 |      +---+---+---+
 |      |  A|  B|  C|
 |      +---+---+---+
 |      |  3|  6|  1|
 |      +---+---+---+
 |
 |      Or positional parameters marked by `?` in the SQL query by SQL literals.
 |
 |      >>> from pyspark.sql.functions import array, lit
 |      >>> spark.sql(
 |      …   "SELECT *, element_at(?, 1) AS C FROM {df} WHERE {df[B]} > ? and ?
< {df[A]}",
 |      …   args=[array(lit(1), lit(2), lit(3)), 5, 2], df=mydf).show()
 |      +---+---+---+
```

```
 |      |  A|  B|  C|
 |      +---+---+---+
 |      |  3|  6|  1|
 |      +---+---+---+
 |
 |  stop(self) -> None
 |      Stop the underlying :class:`SparkContext`.
 |
 |      .. versionadded:: 2.0.0
 |
 |      .. versionchanged:: 3.4.0
 |          Supports Spark Connect.
 |
 |      Examples
 |      --------
 |      >>> spark.stop()  # doctest: +SKIP
 |
 |  streams = <functools.cached_property object>
 |      Returns a :class:`StreamingQueryManager` that allows managing all the
 |      :class:`StreamingQuery` instances active on `this` context.
 |
 |      .. versionadded:: 2.0.0
 |
 |      .. versionchanged:: 3.5.0
 |          Supports Spark Connect.
 |
 |      Notes
 |      -----
 |      This API is evolving.
 |
 |      Returns
 |      -------
 |      :class:`StreamingQueryManager`
 |
 |      Examples
 |      --------
 |      >>> spark.streams
 |      <pyspark…StreamingQueryManager object …>
 |
 |      Get the list of active streaming queries
 |
 |      >>> sq = spark.readStream.format(
 |      …
"rate").load().writeStream.format('memory').queryName('this_query').start()
 |      >>> sqm = spark.streams
 |      >>> [q.name for q in sqm.active]
 |      ['this_query']
 |      >>> sq.stop()
```

```
 |
 |  table(self, tableName: str) -> pyspark.sql.dataframe.DataFrame
 |      Returns the specified table as a :class:`DataFrame`.
 |
 |      .. versionadded:: 2.0.0
 |
 |      .. versionchanged:: 3.4.0
 |          Supports Spark Connect.
 |
 |      Parameters
 |      ----------
 |      tableName : str
 |          the table name to retrieve.
 |
 |      Returns
 |      -------
 |      :class:`DataFrame`
 |
 |      Notes
 |      -----
 |      In Spark Classic, a temporary view referenced in `spark.table` is
resolved immediately,
 |      while in Spark Connect it is lazily analyzed.
 |      So in Spark Connect if a view is dropped, modified or replaced after
`spark.table`, the
 |      execution may fail or generate different results.
 |
 |      Examples
 |      --------
 |      >>> spark.range(5).createOrReplaceTempView("table1")
 |      >>> spark.table("table1").sort("id").show()
 |      +---+
 |      | id|
 |      +---+
 |      |  0|
 |      |  1|
 |      |  2|
 |      |  3|
 |      |  4|
 |      +---+
 |
 |  version = <functools.cached_property object>
 |      The version of Spark on which this application is running.
 |
 |      .. versionadded:: 2.0.0
 |
 |      .. versionchanged:: 3.4.0
 |          Supports Spark Connect.
```

```
 |
 |      Returns
 |      -------
 |      str
 |          the version of Spark in string.
 |
 |      Examples
 |      --------
 |      >>> _ = spark.version
 |
 |  ----------------------------------------------------------------------
 |  Class methods defined here:
 |
 |  active() -> 'SparkSession'
 |      Returns the active or default :class:`SparkSession` for the current
thread, returned by
 |      the builder.
 |
 |      .. versionadded:: 3.5.0
 |
 |      Returns
 |      -------
 |      :class:`SparkSession`
 |          Spark session if an active or default session exists for the current
thread.
 |
 |  getActiveSession() -> Optional[ForwardRef('SparkSession')]
 |      Returns the active :class:`SparkSession` for the current thread,
returned by the builder
 |
 |      .. versionadded:: 3.0.0
 |
 |      .. versionchanged:: 3.5.0
 |          Supports Spark Connect.
 |
 |      Returns
 |      -------
 |      :class:`SparkSession`
 |          Spark session if an active session exists for the current thread
 |
 |      Examples
 |      --------
 |      >>> s = SparkSession.getActiveSession()
 |      >>> df = s.createDataFrame([('Alice', 1)], ['name', 'age'])
 |      >>> df.select("age").show()
 |      +---+
 |      |age|
 |      +---+
```

```
|      |  1|
|      +---+
|
|  ----------------------------------------------------------------------
|  Readonly properties defined here:
|
|  builder
|
|  client
|      Gives access to the Spark Connect client. In normal cases this is not
necessary to be used
|      and only relevant for testing.
|
|      .. versionadded:: 3.4.0
|
|      Returns
|      -------
|      :class:`SparkConnectClient`
|
|      Notes
|      -----
|      This API is unstable, and a developer API. It returns non-API instance
|      :class:`SparkConnectClient`.
|      This is an API dedicated to Spark Connect client only. With regular
Spark Session, it throws
|      an exception.
|
|  dataSource
|      Returns a :class:`DataSourceRegistration` for data source registration.
|
|      .. versionadded:: 4.0.0
|
|      Returns
|      -------
|      :class:`DataSourceRegistration`
|
|      Notes
|      -----
|      This feature is experimental and unstable.
|
|  profile
|      Returns a :class:`Profile` for performance/memory profiling.
|
|      .. versionadded:: 4.0.0
|
|      Returns
|      -------
|      :class:`Profile`
```

```
 |
 |      Notes
 |      -----
 |      Supports Spark Connect.
 |
 |  read
 |      Returns a :class:`DataFrameReader` that can be used to read data
 |      in as a :class:`DataFrame`.
 |
 |      .. versionadded:: 2.0.0
 |
 |      .. versionchanged:: 3.4.0
 |          Supports Spark Connect.
 |
 |      Returns
 |      -------
 |      :class:`DataFrameReader`
 |
 |      Examples
 |      --------
 |      >>> spark.read
 |      <…DataFrameReader object …>
 |
 |      Write a DataFrame into a JSON file and read it back.
 |
 |      >>> import tempfile
 |      >>> with tempfile.TemporaryDirectory(prefix="read") as d:
 |      …     # Write a DataFrame into a JSON file
 |      …     spark.createDataFrame(
 |      …         [{"age": 100, "name": "Hyukjin Kwon"}]
 |      …     ).write.mode("overwrite").format("json").save(d)
 |      …
 |      …     # Read the JSON file as a DataFrame.
 |      …     spark.read.format('json').load(d).show()
 |      +---+------------+
 |      |age|        name|
 |      +---+------------+
 |      |100|Hyukjin Kwon|
 |      +---+------------+
 |
 |  readStream
 |      Returns a :class:`DataStreamReader` that can be used to read data
streams
 |      as a streaming :class:`DataFrame`.
 |
 |      .. versionadded:: 2.0.0
 |
 |      .. versionchanged:: 3.5.0
```

```
 |          Supports Spark Connect.
 |
 |      Notes
 |      -----
 |      This API is evolving.
 |
 |      Returns
 |      -------
 |      :class:`DataStreamReader`
 |
 |      Examples
 |      --------
 |      >>> spark.readStream
 |      <pyspark…DataStreamReader object …>
 |
 |      The example below uses Rate source that generates rows continuously.
 |      After that, we operate a modulo by 3, and then write the stream out to
the console.
 |      The streaming query stops in 3 seconds.
 |
 |      >>> import time
 |      >>> df = spark.readStream.format("rate").load()
 |      >>> df = df.selectExpr("value % 3 as v")
 |      >>> q = df.writeStream.format("console").start()
 |      >>> time.sleep(3)
 |      >>> q.stop()
 |
 |  sparkContext
 |      Returns the underlying :class:`SparkContext`.
 |
 |      .. versionadded:: 2.0.0
 |
 |      Returns
 |      -------
 |      :class:`SparkContext`
 |
 |      Examples
 |      --------
 |      >>> spark.sparkContext
 |      <SparkContext master=… appName=…>
 |
 |      Create an RDD from the Spark context
 |
 |      >>> rdd = spark.sparkContext.parallelize([1, 2, 3])
 |      >>> rdd.collect()
 |      [1, 2, 3]
 |
 |  tvf
```

```
 |      Returns a :class:`tvf.TableValuedFunction` that can be used to call a
table-valued function
 |      (TVF).
 |
 |      .. versionadded:: 4.0.0
 |
 |      Notes
 |      -----
 |      This API is evolving.
 |
 |      Returns
 |      -------
 |      :class:`tvf.TableValuedFunction`
 |
 |      Examples
 |      --------
 |      >>> spark.tvf
 |      <pyspark…TableValuedFunction object …>
 |
 |      >>> import pyspark.sql.functions as sf
 |      >>> spark.tvf.explode(sf.array(sf.lit(1), sf.lit(2), sf.lit(3))).show()
 |      +---+
 |      |col|
 |      +---+
 |      |  1|
 |      |  2|
 |      |  3|
 |      +---+
 |
 |  udf
 |      Returns a :class:`UDFRegistration` for UDF registration.
 |
 |      .. versionadded:: 2.0.0
 |
 |      .. versionchanged:: 3.4.0
 |          Supports Spark Connect.
 |
 |      Returns
 |      -------
 |      :class:`UDFRegistration`
 |
 |      Examples
 |      --------
 |      Register a Python UDF, and use it in SQL.
 |
 |      >>> strlen = spark.udf.register("strlen", lambda x: len(x))
 |      >>> spark.sql("SELECT strlen('test')").show()
 |      +------------+
```

```
 |      |strlen(test)|
 |      +------------+
 |      |           4|
 |      +------------+
 |
 |  udtf
 |      Returns a :class:`UDTFRegistration` for UDTF registration.
 |
 |      .. versionadded:: 3.5.0
 |
 |      Returns
 |      -------
 |      :class:`UDTFRegistration`
 |
 |      Notes
 |      -----
 |      Supports Spark Connect.
 |
 |  ----------------------------------------------------------------------
 |  Data and other attributes defined here:
 |
 |  Builder = <class 'pyspark.sql.session.SparkSession.Builder'>
 |      Builder for :class:`SparkSession`.
 |
 |
 |  __annotations__ = {'_activeSession': typing.ClassVar[typing.Optional[F…
 |
 |  ----------------------------------------------------------------------
 |  Data descriptors inherited from
pyspark.sql.pandas.conversion.SparkConversionMixin:
 |
 |  __dict__
 |      dictionary for instance variables
 |
 |  __weakref__
 |      list of weak references to the object
```

[13]:
```
# Help can be used on any Python object
help(map)
help(testmti850.Test)
```

```
Help on class map in module builtins:

class map(object)
 |  map(function, iterable, /, *iterables)
 |
 |  Make an iterator that computes the function using arguments from
```

```
 |  each of the iterables.  Stops when the shortest iterable is exhausted.
 |
 |  Methods defined here:
 |
 |  __getattribute__(self, name, /)
 |      Return getattr(self, name).
 |
 |  __iter__(self, /)
 |      Implement iter(self).
 |
 |  __next__(self, /)
 |      Implement next(self).
 |
 |  __reduce__(self, /)
 |      Return state information for pickling.
 |
 |  ----------------------------------------------------------------------
 |  Static methods defined here:
 |
 |  __new__(*args, **kwargs)
 |      Create and return a new object.  See help(type) for accurate signature.

Help on class Test in module testmti850:

class Test(builtins.object)
 |  Class methods defined here:
 |
 |  assertEquals(var, val, msg='')
 |
 |  assertEqualsHashed(var, hashed_val, msg='')
 |
 |  assertTrue(result, msg='')
 |
 |  printStats()
 |
 |  setFailFast()
 |
 |  setPrivateMode()
 |
 |  ----------------------------------------------------------------------
 |  Data descriptors defined here:
 |
 |  __dict__
 |      dictionary for instance variables
 |
 |  __weakref__
 |      list of weak references to the object
 |
```

```
|  ----------------------------------------------------------------------
|  Data and other attributes defined here:
|
|  failFast = False
|
|  numTests = 0
|
|  passed = 0
|
|  private = False
```

1.3 Part 2: Exploratory Data Analysis

Let's begin looking at our data.

For this assignment, we will use a data set from NASA Kennedy Space Center web server in Florida.

The full data set is freely available at ftp://ita.ee.lbl.gov/html/contrib/NASA-HTTP.html, and it contains all HTTP requests for two months.

We are using a subset that only contains several days' worth of requests. To download the log and put it into HDFS, run the following commands in a terminal:

```
wget ftp://ita.ee.lbl.gov/traces/NASA_access_log_Aug95.gz
gunzip NASA_access_log_Aug95.gz
hdfs dfs -put NASA_access_log_Aug95 /NASA_access_log_Aug95.txt
```

1.3.1 (2a) Loading the log file

Now that we have the path to the file, let's load it into a DataFrame. We'll do this in steps. First, we'll use `spark.read.text()` to read the text file. This will produce a DataFrame with a single string column called `value`.

[15]:
```
log_filename = "hdfs://localhost:9000/NASA_access_log_Aug95.txt"
base_df = spark.read.text(log_filename)

# Let's look at the schema
base_df.printSchema()
```

```
root
 |-- value: string (nullable = true)
```

Let's take a look at some of the data.

[16]:
```
base_df.show(truncate=False)
```

```
+-----------------------------------------------------------------------
----------------------------------------------+
|value
|
```

```
+--------------------------------------------------------------------------------
-----------------------------------------------+
|in24.inetnebr.com - - [01/Aug/1995:00:00:01 -0400] "GET
/shuttle/missions/sts-68/news/sts-68-mcc-05.txt HTTP/1.0" 200 1839     |
|uplherc.upl.com - - [01/Aug/1995:00:00:07 -0400] "GET / HTTP/1.0" 304 0
|
|uplherc.upl.com - - [01/Aug/1995:00:00:08 -0400] "GET /images/ksclogo-
medium.gif HTTP/1.0" 304 0                                    |
|uplherc.upl.com - - [01/Aug/1995:00:00:08 -0400] "GET /images/MOSAIC-
logosmall.gif HTTP/1.0" 304 0                                  |
|uplherc.upl.com - - [01/Aug/1995:00:00:08 -0400] "GET /images/USA-logosmall.gif
HTTP/1.0" 304 0                                  |
|ix-esc-ca2-07.ix.netcom.com - - [01/Aug/1995:00:00:09 -0400] "GET
/images/launch-logo.gif HTTP/1.0" 200 1713                    |
|uplherc.upl.com - - [01/Aug/1995:00:00:10 -0400] "GET /images/WORLD-
logosmall.gif HTTP/1.0" 304 0                                   |
|slppp6.intermind.net - - [01/Aug/1995:00:00:10 -0400] "GET
/history/skylab/skylab.html HTTP/1.0" 200 1687                       |
|piweba4y.prodigy.com - - [01/Aug/1995:00:00:10 -0400] "GET
/images/launchmedium.gif HTTP/1.0" 200 11853                         |
|slppp6.intermind.net - - [01/Aug/1995:00:00:11 -0400] "GET
/history/skylab/skylab-small.gif HTTP/1.0" 200 9202                  |
|slppp6.intermind.net - - [01/Aug/1995:00:00:12 -0400] "GET
/images/ksclogosmall.gif HTTP/1.0" 200 3635                          |
|ix-esc-ca2-07.ix.netcom.com - - [01/Aug/1995:00:00:12 -0400] "GET
/history/apollo/images/apollo-logo1.gif HTTP/1.0" 200 1173   |
|slppp6.intermind.net - - [01/Aug/1995:00:00:13 -0400] "GET
/history/apollo/images/apollo-logo.gif HTTP/1.0" 200 3047           |
|uplherc.upl.com - - [01/Aug/1995:00:00:14 -0400] "GET /images/NASA-
logosmall.gif HTTP/1.0" 304 0                                    |
|133.43.96.45 - - [01/Aug/1995:00:00:16 -0400] "GET
/shuttle/missions/sts-69/mission-sts-69.html HTTP/1.0" 200 10566            |
|kgtyk4.kj.yamagata-u.ac.jp - - [01/Aug/1995:00:00:17 -0400] "GET / HTTP/1.0"
200 7280                                          |
|kgtyk4.kj.yamagata-u.ac.jp - - [01/Aug/1995:00:00:18 -0400] "GET
/images/ksclogo-medium.gif HTTP/1.0" 200 5866                 |
|d0ucr6.fnal.gov - - [01/Aug/1995:00:00:19 -0400] "GET
/history/apollo/apollo-16/apollo-16.html HTTP/1.0" 200 2743              |
|ix-esc-ca2-07.ix.netcom.com - - [01/Aug/1995:00:00:19 -0400] "GET
/shuttle/resources/orbiters/discovery.html HTTP/1.0" 200 6849|
|d0ucr6.fnal.gov - - [01/Aug/1995:00:00:20 -0400] "GET
/history/apollo/apollo-16/apollo-16-patch-small.gif HTTP/1.0" 200 14897  |
+--------------------------------------------------------------------------------
-----------------------------------------------+
only showing top 20 rows
```

1.3.2 (2b) Parsing the log file

If you're familiar with web servers at all, you'll recognize that this is in Common Log Format. The fields are:

remotehost rfc931 authuser [date] "request" status bytes

field	meaning
remotehost	Remote hostname (or IP number if DNS hostname is not available).
rfc931	The remote logname of the user. We don't really care about this field.
authuser	The username of the remote user, as authenticated by the HTTP server.
[date]	The date and time of the request.
"request"	The request, exactly as it came from the browser or client.
status	The HTTP status code the server sent back to the client.
bytes	The number of bytes (`Content-Length`) transferred to the client.

Next, we have to parse it into individual columns. We'll use the special built-in regexp_extract() function to do the parsing. This function matches a column against a regular expression with one or more capture groups and allows you to extract one of the matched groups. We'll use one regular expression for each field we wish to extract.

If you can't read these regular expressions, don't worry. Trust us: They work. If you find regular expressions confusing (and they certainly *can* be), and you want to learn more about them, start with the RegexOne web site. You might also find *Regular Expressions Cookbook*, by Jan Goyvaerts and Steven Levithan, to be helpful.

Some people, when confronted with a problem, think "I know, I'll use regular expressions." Now they have two problems. (attributed to Jamie Zawinski)

[17]:
```
from pyspark.sql.functions import split, regexp_extract

split_df = base_df.select(regexp_extract('value', r'^([^\s]+\s)', 1).
 ↪alias('host'),
                          regexp_extract('value', r'^.*\[(\d\d/\w{3}/\d{4}:
 ↪\d{2}:\d{2}:\d{2} -\d{4})]', 1).alias('timestamp'),
                          regexp_extract('value', r'^.*"\w+\s+([^\s]+)\s+HTTP.
 ↪*"', 1).alias('path'),
                          regexp_extract('value', r'^.*"\s+([^\s]+)', 1).
 ↪alias('status'),
                          regexp_extract('value', r'^.*\s+(\d+)$', 1).
 ↪alias('content_size'))

split_df.show(truncate=False)
```

```
+--------------------------+-------------------------+----------------------
--------------------------+------+------------+
|host                      |timestamp                |path
|status|content_size|
+--------------------------+-------------------------+----------------------
```

```
---------------------------+------+-----------+
|in24.inetnebr.com           |01/Aug/1995:00:00:01
-0400|/shuttle/missions/sts-68/news/sts-68-mcc-05.txt    |200   |1839        |
|uplherc.upl.com             |01/Aug/1995:00:00:07 -0400|/
|304   |0           |
|uplherc.upl.com             |01/Aug/1995:00:00:08 -0400|/images/ksclogo-
medium.gif                        |304   |0           |
|uplherc.upl.com             |01/Aug/1995:00:00:08 -0400|/images/MOSAIC-
logosmall.gif                       |304   |0           |
|uplherc.upl.com             |01/Aug/1995:00:00:08 -0400|/images/USA-
logosmall.gif                          |304   |0           |
|ix-esc-ca2-07.ix.netcom.com |01/Aug/1995:00:00:09 -0400|/images/launch-logo.gif
|200   |1713        |
|uplherc.upl.com             |01/Aug/1995:00:00:10 -0400|/images/WORLD-
logosmall.gif                         |304   |0           |
|slppp6.intermind.net        |01/Aug/1995:00:00:10
-0400|/history/skylab/skylab.html                         |200   |1687        |
|piweba4y.prodigy.com        |01/Aug/1995:00:00:10
-0400|/images/launchmedium.gif                            |200   |11853       |
|slppp6.intermind.net        |01/Aug/1995:00:00:11 -0400|/history/skylab/skylab-
small.gif                    |200   |9202        |
|slppp6.intermind.net        |01/Aug/1995:00:00:12
-0400|/images/ksclogosmall.gif                            |200   |3635        |
|ix-esc-ca2-07.ix.netcom.com |01/Aug/1995:00:00:12
-0400|/history/apollo/images/apollo-logo1.gif             |200   |1173        |
|slppp6.intermind.net        |01/Aug/1995:00:00:13
-0400|/history/apollo/images/apollo-logo.gif              |200   |3047        |
|uplherc.upl.com             |01/Aug/1995:00:00:14 -0400|/images/NASA-
logosmall.gif                          |304   |0           |
|133.43.96.45                |01/Aug/1995:00:00:16
-0400|/shuttle/missions/sts-69/mission-sts-69.html        |200   |10566       |
|kgtyk4.kj.yamagata-u.ac.jp  |01/Aug/1995:00:00:17 -0400|/
|200   |7280        |
|kgtyk4.kj.yamagata-u.ac.jp  |01/Aug/1995:00:00:18 -0400|/images/ksclogo-
medium.gif                        |200   |5866        |
|d0ucr6.fnal.gov             |01/Aug/1995:00:00:19
-0400|/history/apollo/apollo-16/apollo-16.html            |200   |2743        |
|ix-esc-ca2-07.ix.netcom.com |01/Aug/1995:00:00:19
-0400|/shuttle/resources/orbiters/discovery.html          |200   |6849        |
|d0ucr6.fnal.gov             |01/Aug/1995:00:00:20
-0400|/history/apollo/apollo-16/apollo-16-patch-small.gif|200   |14897       |
+---------------------------+-------------------------+----------------------
---------------------------+------+-----------+
only showing top 20 rows
```

1.3.3 (2c) Data Cleaning

Let's see how well our parsing logic worked. First, let's verify that there are no null rows in the original data set.

[18]:
```
base_df.filter(base_df['value'].isNull()).count()
```

[18]:
```
0
```

If our parsing worked properly, we'll have no rows with null column values. Let's check.

[20]:
```
from pyspark.sql.functions import col, expr

split_df = (split_df
    .withColumn("status", expr("try_cast(status AS INT)"))
    .withColumn("content_size", expr("try_cast(content_size AS INT)"))
)
```

[21]:
```
bad_rows_df = split_df.filter(split_df['host'].isNull() |
                              split_df['timestamp'].isNull() |
                              split_df['path'].isNull() |
                              split_df['status'].isNull() |
                             split_df['content_size'].isNull())

bad_rows_df.count()
```

[21]:
```
14178
```

Not good. We have some null values. Something went wrong. Which columns are affected?

(Note: This approach is adapted from an excellent answer on StackOverflow.)

[22]:
```
from pyspark.sql.functions import col, sum

def count_null(col_name):
    return sum(col(col_name).isNull().cast('integer')).alias(col_name)

# Build up a list of column expressions, one per column.
#
# This could be done in one line with a Python list comprehension, but we're␣
 ↪keeping
# it simple for those who don't know Python very well.
exprs = []
for col_name in split_df.columns:
    exprs.append(count_null(col_name))

# Run the aggregation. The *exprs converts the list of expressions into
```

```
# variable function arguments.
split_df.agg(*exprs).show() # aggregate on the entire DataFrame (split_df)␣
 ↪without groups using the column expressions
```

```
[Stage 10:============================>                            (1 + 1) / 2]

+----+---------+----+------+------------+
|host|timestamp|path|status|content_size|
+----+---------+----+------+------------+
|   0|        0|   0|     0|       14178|
+----+---------+----+------+------------+
```

Okay, they're all in the `content_size` column. Let's see if we can figure out what's wrong. Our original parsing regular expression for that column was:

```
regexp_extract('value', r'^.*\s+(\d+)$', 1).cast('integer').alias('content_size')
```

The `\d+` selects one or more digits at the end of the input line. Is it possible there are lines without a valid content size? Or is there something wrong with our regular expression? Let's see if there are any lines that do not end with one or more digits.

Note: In the expression below, `~` means "not".

[23]:
```
bad_content_size_df = base_df.filter(~ base_df['value'].rlike(r'\d+$')) #␣
 ↪base_df is the dataframe with the raw log (before split)

bad_content_size_df.count()
```

[23]:
```
14178
```

That's it! The count matches the number of rows in `bad_rows_df` exactly.

Let's take a look at some of the bad column values. Since it's possible that the rows end in extra white space, we'll tack a marker character onto the end of each line, to make it easier to see trailing white space.

[24]:
```
from pyspark.sql.functions import lit, concat

bad_content_size_df.select(concat(bad_content_size_df['value'], lit('*'))).
 ↪show(truncate=False)
```

```
+-----------------------------------------------------------------------------
-----------------------------------------+
|concat(value, *)
|
+-----------------------------------------------------------------------------
-----------------------------------------+
|gw1.att.com - - [01/Aug/1995:00:03:53 -0400] "GET /shuttle/missions/sts-73/news
```

```
HTTP/1.0" 302 -*                                |
|js002.cc.utsunomiya-u.ac.jp - - [01/Aug/1995:00:07:33 -0400] "GET
/shuttle/resources/orbiters/discovery.gif HTTP/1.0" 404 -*|
|tia1.eskimo.com - - [01/Aug/1995:00:28:41 -0400] "GET /pub/winvn/release.txt
HTTP/1.0" 404 -*                                  |
|itws.info.eng.niigata-u.ac.jp - - [01/Aug/1995:00:38:01 -0400] "GET
/ksc.html/facts/about_ksc.html HTTP/1.0" 403 -*          |
|grimnet23.idirect.com - - [01/Aug/1995:00:50:12 -0400] "GET
/www/software/winvn/winvn.html HTTP/1.0" 404 -*                  |
|miriworld.its.unimelb.edu.au - - [01/Aug/1995:01:04:54 -0400] "GET
/history/history.htm HTTP/1.0" 404 -*                     |
|ras38.srv.net - - [01/Aug/1995:01:05:14 -0400] "GET /elv/DELTA/uncons.htm
HTTP/1.0" 404 -*                                    |
|cs1-06.leh.ptd.net - - [01/Aug/1995:01:17:38 -0400] "GET /sts-71/launch/" 404
-*                                                 |
|www-b2.proxy.aol.com - - [01/Aug/1995:01:22:07 -0400] "GET /shuttle/countdown
HTTP/1.0" 302 -*                                   |
|maui56.maui.net - - [01/Aug/1995:01:31:56 -0400] "GET /shuttle HTTP/1.0" 302 -*
|
|dialip-24.athenet.net - - [01/Aug/1995:01:33:02 -0400] "GET
/history/apollo/apollo-13.html HTTP/1.0" 404 -*                  |
|h96-158.ccnet.com - - [01/Aug/1995:01:35:50 -0400] "GET
/history/apollo/a-001/a-001-patch-small.gif HTTP/1.0" 404 -*        |
|h96-158.ccnet.com - - [01/Aug/1995:01:36:23 -0400] "GET
/history/apollo/a-001/movies/ HTTP/1.0" 404 -*                      |
|h96-158.ccnet.com - - [01/Aug/1995:01:36:30 -0400] "GET
/history/apollo/a-001/a-001-patch-small.gif HTTP/1.0" 404 -*        |
|h96-158.ccnet.com - - [01/Aug/1995:01:36:38 -0400] "GET
/history/apollo/a-001/movies/ HTTP/1.0" 404 -*                      |
|h96-158.ccnet.com - - [01/Aug/1995:01:36:42 -0400] "GET
/history/apollo/a-001/a-001-patch-small.gif HTTP/1.0" 404 -*        |
|h96-158.ccnet.com - - [01/Aug/1995:01:36:44 -0400] "GET
/history/apollo/a-001/images/ HTTP/1.0" 404 -*                      |
|h96-158.ccnet.com - - [01/Aug/1995:01:36:47 -0400] "GET
/history/apollo/a-001/a-001-patch-small.gif HTTP/1.0" 404 -*        |
|h96-158.ccnet.com - - [01/Aug/1995:01:37:04 -0400] "GET
/history/apollo/a-004/a-004-patch-small.gif HTTP/1.0" 404 -*        |
|h96-158.ccnet.com - - [01/Aug/1995:01:37:05 -0400] "GET
/history/apollo/a-004/movies/ HTTP/1.0" 404 -*                      |
+-------------------------------------------------------------------------------
-----------------------------------------+
only showing top 20 rows
```

Ah. The bad rows correspond to error results, where no content was sent back and the server emitted a "-" for the `content_size` field. Since we don't want to discard those rows from our analysis, let's map them to 0.

1.3.4 (2d) Fix the rows with null content_size

The easiest solution is to replace the null values in `split_df` with 0. The DataFrame API provides a set of functions and fields specifically designed for working with null values, among them:

- fillna(), which fills null values with specified non-null values.
- na, which returns a DataFrameNaFunctions object with many functions for operating on null columns.

We'll use `fillna()`, because it's simple. There are several ways to invoke this function. The easiest is just to replace *all* null columns with known values. But, for safety, it's better to pass a Python dictionary containing (column_name, value) mappings. That's what we'll do.

[25]:
```
# Replace all null content_size values with 0.

cleaned_df = split_df.fillna({'content_size': 0}) # in fact, fillna() is an␣
 ↪alias for na.fill()
```

Now, let us ensure that there are no nulls left.

[26]:
```
exprs = []
for col_name in cleaned_df.columns:
  exprs.append(count_null(col_name))

cleaned_df.agg(*exprs).show()
```

```
[Stage 17:==============================>                          (1 + 1) / 2]

+----+---------+----+------+------------+
|host|timestamp|path|status|content_size|
+----+---------+----+------+------------+
|   0|        0|   0|     0|           0|
+----+---------+----+------+------------+
```

1.3.5 (2e) Parsing the timestamp.

Okay, now that we have a clean, parsed DataFrame, we have to parse the timestamp field into an actual timestamp. The Common Log Format time is somewhat non-standard. A User-Defined Function (UDF) is the most straightforward way to parse it.

[27]:
```
month_map = {
  'Jan': 1, 'Feb': 2, 'Mar':3, 'Apr':4, 'May':5, 'Jun':6, 'Jul':7,
  'Aug':8,  'Sep': 9, 'Oct':10, 'Nov': 11, 'Dec': 12
}

def parse_clf_time(s):
    """ Convert Common Log time format into a Python datetime object
    Args:
```

```
        s (str): date and time in Apache time format [dd/mmm/yyyy:hh:mm:ss (+/
 ↪-)zzzz]
    Returns:
        a string suitable for passing to CAST('timestamp')
    """
    # NOTE: We're ignoring time zone here. In a production application, you'd␣
 ↪want to handle that.
    return "{0:04d}-{1:02d}-{2:02d} {3:02d}:{4:02d}:{5:02d}".format(
      int(s[7:11]),
      month_map[s[3:6]],
      int(s[0:2]),
      int(s[12:14]),
      int(s[15:17]),
      int(s[18:20])
    )

u_parse_time = spark.udf.register('parse_clf_time', parse_clf_time) # register␣
 ↪the UDF

# sequence of operations: select all columns, add a new one (time), and remove␣
 ↪the timestamp column
logs_df = cleaned_df.select('*', u_parse_time(cleaned_df['timestamp']).
 ↪cast('timestamp').alias('time')).drop('timestamp')

total_log_entries = logs_df.count() # keep the total of log entries for future␣
 ↪operations

logs_df.show()
```

```
+--------------------+--------------------+------+-----------+----------------
--+
|                host|                path|status|content_size|
time|
+--------------------+--------------------+------+-----------+----------------
--+
|  in24.inetnebr.com |/shuttle/missions…|   200|        1839|1995-08-01
00:00:01|
|    uplherc.upl.com |                   /|   304|           0|1995-08-01
00:00:07|
|    uplherc.upl.com |/images/ksclogo-m…|   304|           0|1995-08-01
00:00:08|
|    uplherc.upl.com |/images/MOSAIC-lo…|   304|           0|1995-08-01
00:00:08|
|    uplherc.upl.com |/images/USA-logos…|   304|           0|1995-08-01
00:00:08|
|ix-esc-ca2-07.ix…|/images/launch-lo…|   200|        1713|1995-08-01
00:00:09|
```

```
|    uplherc.upl.com |/images/WORLD-log…|   304|           0|1995-08-01
00:00:10|
|slppp6.intermind…|/history/skylab/s…|   200|        1687|1995-08-01
00:00:10|
|piweba4y.prodigy…|/images/launchmed…|   200|       11853|1995-08-01
00:00:10|
|slppp6.intermind…|/history/skylab/s…|   200|        9202|1995-08-01
00:00:11|
|slppp6.intermind…|/images/ksclogosm…|   200|        3635|1995-08-01
00:00:12|
|ix-esc-ca2-07.ix…|/history/apollo/i…|   200|        1173|1995-08-01
00:00:12|
|slppp6.intermind…|/history/apollo/i…|   200|        3047|1995-08-01
00:00:13|
|    uplherc.upl.com |/images/NASA-logo…|   304|           0|1995-08-01
00:00:14|
|       133.43.96.45 |/shuttle/missions…|   200|       10566|1995-08-01
00:00:16|
|kgtyk4.kj.yamagat…|                   /|   200|        7280|1995-08-01
00:00:17|
|kgtyk4.kj.yamagat…|/images/ksclogo-m…|   200|        5866|1995-08-01
00:00:18|
|    d0ucr6.fnal.gov |/history/apollo/a…|   200|        2743|1995-08-01
00:00:19|
|ix-esc-ca2-07.ix…|/shuttle/resource…|   200|        6849|1995-08-01
00:00:19|
|    d0ucr6.fnal.gov |/history/apollo/a…|   200|       14897|1995-08-01
00:00:20|
+--------------------+--------------------+------+------------+----------------
--+
only showing top 20 rows
```

[28]:
```
logs_df.printSchema()
```

```
root
 |-- host: string (nullable = true)
 |-- path: string (nullable = true)
 |-- status: integer (nullable = true)
 |-- content_size: integer (nullable = false)
 |-- time: timestamp (nullable = true)
```

[29]:
```
display(logs_df)
```

```
DataFrame[host: string, path: string, status: int, content_size: int, time:␣
↪timestamp]
```

Let's cache `logs_df`. We're going to be using it quite a bit from here forward.

[30]:
```
logs_df.cache()
```

[30]:
```
DataFrame[host: string, path: string, status: int, content_size: int, time:
timestamp]
```

1.4 Part 3: Analysis Walk-Through on the Web Server Log File

Now that we have a DataFrame containing the parsed log file as a set of Row objects, we can perform various analyses.

1.4.1 (3a) Example: Content Size Statistics

Let's compute some statistics about the sizes of content being returned by the web server. In particular, we'd like to know what are the average, minimum, and maximum content sizes.

We can compute the statistics by calling `.describe()` on the `content_size` column of `logs_df`. The `.describe()` function returns the count, mean, stddev, min, and max of a given column.

[31]:
```
# Calculate statistics based on the content size.

content_size_summary_df = logs_df.describe(['content_size'])

content_size_summary_df.show()
```

```
[Stage 24:==============================>                          (1 + 1) / 2]

+-------+------------------+
|summary|      content_size|
+-------+------------------+
|  count|           1569898|
|   mean|17089.225812122826|
| stddev|  67954.7639215694|
|    min|                 0|
|    max|           3421948|
+-------+------------------+
```

Alternatively, we can use SQL to directly calculate these statistics. You can explore the many useful functions within the `pyspark.sql.functions` module in the documentation.

After we apply the `.agg()` function, we call `.first()` to extract the first value, which is equivalent to `.take(1)[0]`.

[32]:
```
import pyspark.sql.functions as sqlFunctions

content_size_stats =  (logs_df
                       .agg(sqlFunctions.min(logs_df['content_size']),
                            sqlFunctions.avg(logs_df['content_size']),
                            sqlFunctions.max(logs_df['content_size'])))
```

```
                        .first())

print('Using SQL functions:')
print('Content Size Avg: {1:,.2f}; Min: {0:.2f}; Max: {2:,.0f}'.
  ↪format(*content_size_stats))
```

```
Using SQL functions:
Content Size Avg: 17,089.23; Min: 0.00; Max: 3,421,948
```

1.4.2 (3b) Example: HTTP Status Analysis

Next, let's look at the status values that appear in the log. We want to know which status values appear in the data and how many times. We again start with `logs_df`, then group by the `status` column, apply the `.count()` aggregation function, and sort by the `status` column.

[33]:
```
status_to_count_df =(logs_df
                     .groupBy('status')
                     .count()
                     .sort('status')
                     .cache())

status_to_count_length = status_to_count_df.count()

print ('Found %d response codes' % status_to_count_length)

status_to_count_df.show()

assert status_to_count_length == 8
assert status_to_count_df.take(100) == [(200, 1398988), (302, 26497), (304,␣
  ↪134146), (400, 10), (403, 171), (404, 10056), (500, 3), (501, 27)]
```

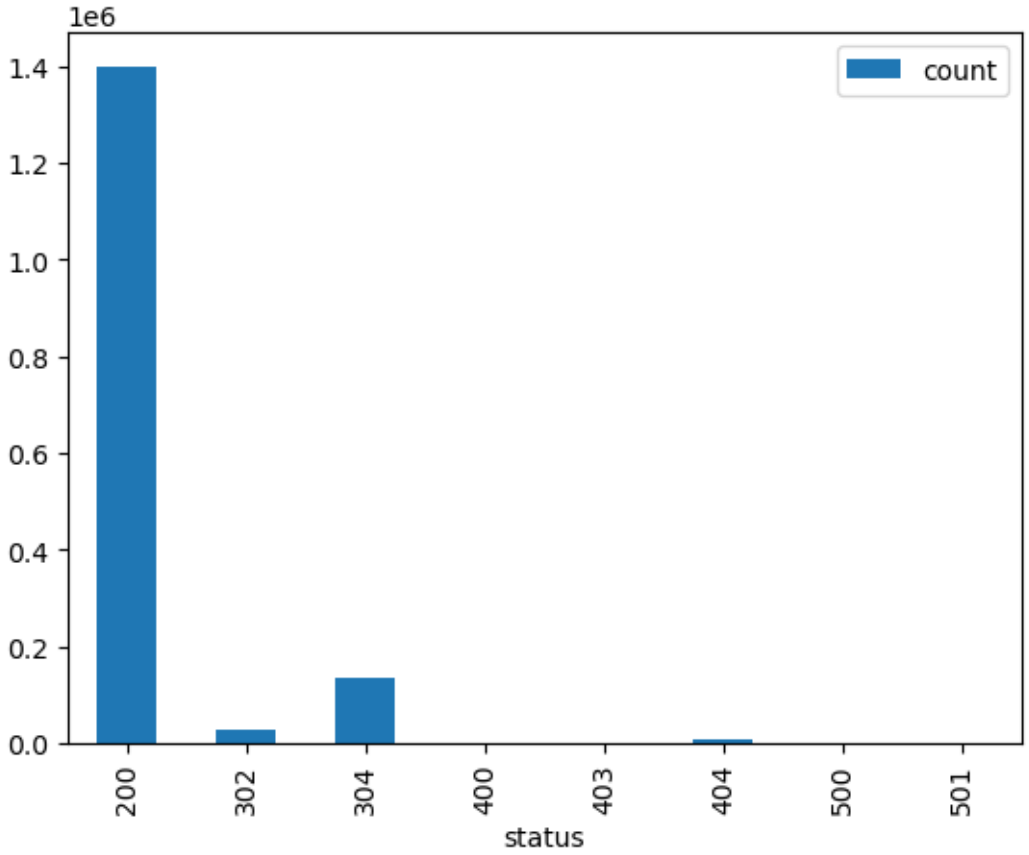
```
Found 8 response codes
+------+-------+
|status|  count|
+------+-------+
|   200|1398988|
|   302|  26497|
|   304| 134146|
|   400|     10|
|   403|    171|
|   404|  10056|
|   500|      3|
|   501|     27|
+------+-------+
```

1.4.3 (3c) Example: Status Graphing

Now, let's visualize the results from the last example. We can convert the original pyspark dataframe into a Pandas dataframe and then use the plot functionality. See the next cell for an example.

[34]:
```
status_to_count_df_pd = status_to_count_df.toPandas()

status_to_count_df_pd.plot.bar(x='status', y='count')
```

[34]:
```
<Axes: xlabel='status'>
```

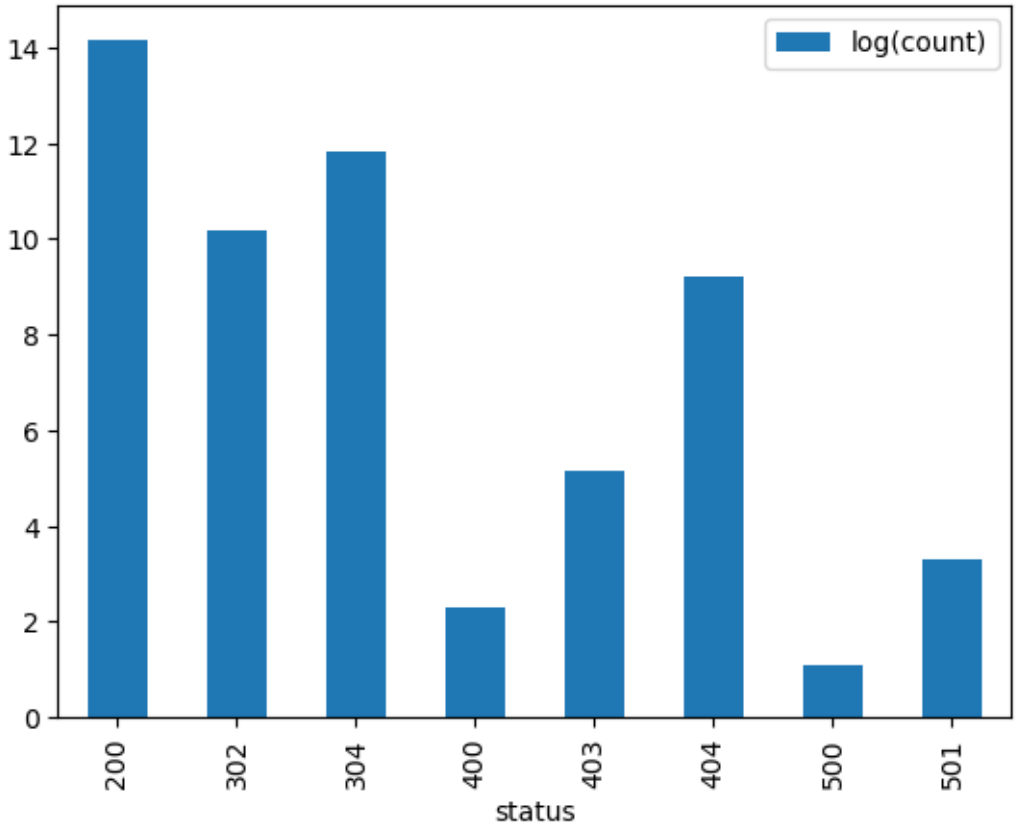


You can see that this is not a very effective plot. Due to the large number of '200' codes, it is very hard to see the relative number of the others. We can alleviate this by taking the logarithm of the count, adding that as a column to our DataFrame and displaying the result.

[35]:
```
log_status_to_count_df = status_to_count_df.withColumn('log(count)',␣
  ↪sqlFunctions.log(status_to_count_df['count']))

log_status_to_count_df_pd = log_status_to_count_df.toPandas()
```

```
log_status_to_count_df_pd.plot.bar(x='status', y='log(count)')
```

[35]: <Axes: xlabel='status'>



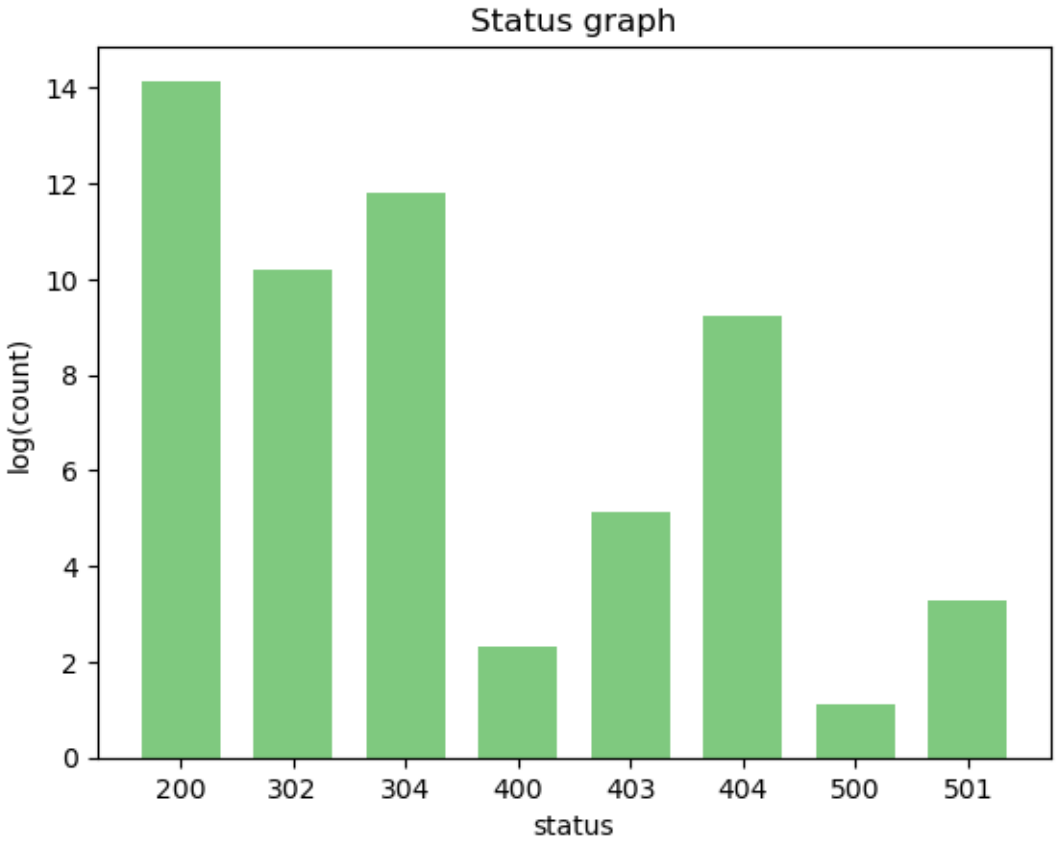
While this graph is an improvement, we might want to make more adjustments. The `matplotlib` library can give us more control in our plot. In this case, we're essentially just reproducing the Pandas graph using `matplotlib`. However, `matplotlib` exposes far more controls than the Pandas graph, allowing you to change colors, label the axes, and more.

We're using the "Set1" color map. See the list of Qualitative Color Maps at http://matplotlib.org/examples/color/colormaps_reference.html for more details. Feel free to change the color map to a different one, like "Accent".

[37]:
```
import numpy as np
import matplotlib.pyplot as plt

data = log_status_to_count_df.drop('count').collect()
x, y = zip(*data) # split status (x) and count (y)
index = np.arange(len(x))
bar_width = 0.7
colorMap = 'Accent'
```

```
cmap = plt.cm.get_cmap(colorMap)

fig, ax = plt.subplots(nrows=1, ncols=1)
plt.bar(index, y, width=bar_width, color=cmap(0))
plt.xticks(ticks=index, labels=x) # set markers (ticks) and the respective␣
 ↪labels in the x-axis
plt.yticks(ticks=np.arange(0, 15, 2))
plt.xlabel('status')
plt.ylabel('log(count)')
plt.title('Status graph')
plt.show()
```

```
/tmp/ipykernel_4389/135583237.py:9: MatplotlibDeprecationWarning: The get_cmap
function was deprecated in Matplotlib 3.7 and will be removed in 3.11. Use
``matplotlib.colormaps[name]`` or ``matplotlib.colormaps.get_cmap()`` or
``pyplot.get_cmap()`` instead.
  cmap = plt.cm.get_cmap(colorMap)
```



1.4.4 (3d) Example: Frequent Hosts

Let's look at hosts that have accessed the server frequently (e.g., more than ten times). As with the response code analysis in (3b), we create a new DataFrame by grouping `logs_df` by the 'host' column and aggregating by count.

We then filter the result based on the count of accesses by each host being greater than ten. Then, we select the 'host' column and show 20 elements from the result.

[38]:
```
# Any hosts that has accessed the server more than 10 times.
host_sum_df =(logs_df
              .groupBy('host')
              .count())

host_more_than_10_df = (host_sum_df
                        .filter(host_sum_df['count'] > 10)
                        .select(host_sum_df['host'])) # select only the host␣
 ↪column

print ('Any 20 hosts that have accessed more then 10 times:\n')

host_more_than_10_df.show(truncate=False)
```

```
Any 20 hosts that have accessed more then 10 times:

[Stage 64:>                                                         (0 + 1) / 2]

+--------------------------------+
|host                            |
+--------------------------------+
|prakinf2.prakinf.tu-ilmenau.de |
|alpha2.csd.uwm.edu              |
|cjc07992.slip.digex.net         |
|n1377004.ksc.nasa.gov           |
|163.205.2.134                   |
|huge.oso.chalmers.se            |
|163.205.44.27                   |
|shark.ksc.nasa.gov              |
|etc5.etechcorp.com              |
|dd07-029.compuserve.com         |
|131.182.101.161                 |
|134.95.100.201                  |
|vab08.larc.nasa.gov             |
|ip11.iac.net                    |
|ad11-012.compuserve.com         |
|ad053.du.pipex.com              |
|204.184.6.19                    |
|p8.denver1.dialup.csn.net       |
|gate2.gdc.com                   |
```

```
|alcott.acsu.buffalo.edu        |
+-------------------------------+
only showing top 20 rows
```

1.4.5 (3e) Example: Visualizing Paths

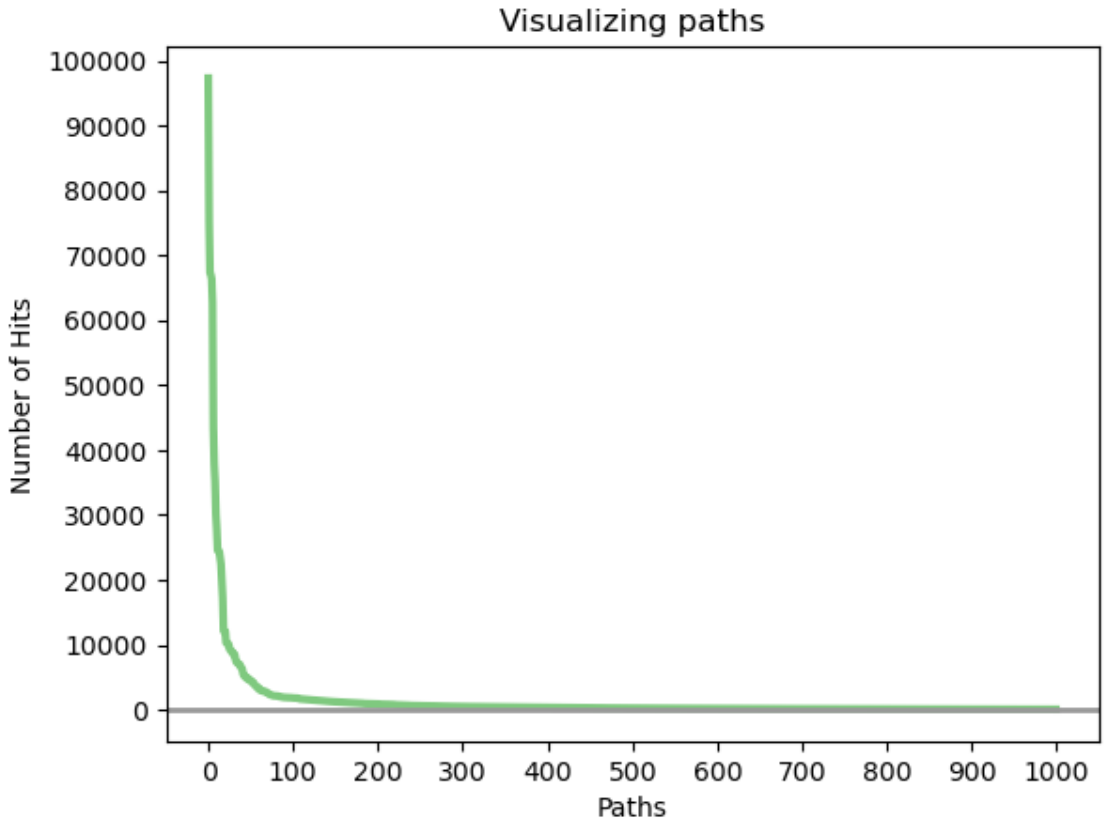
Now, let's visualize the number of hits to paths (URIs) in the log. To perform this task, we start with our `logs_df` and group by the `path` column, aggregate by count, and sort in descending order.

Next we visualize the results using `matplotlib`. We extract the paths and the counts, and unpack the resulting list of `Rows` using a `map` function and `lambda` expression.

[39]:
```
paths_df = (logs_df
            .groupBy('path')
            .count()
            .sort('count', ascending=False)) # two-column dataframe

paths_counts = (paths_df
                .select('path', 'count') # in this case, it is similar to '*'
                .rdd.map(lambda r: (r[0], r[1]))
                .collect())

paths, counts = zip(*paths_counts)

colorMap = 'Accent'
cmap = plt.cm.get_cmap(colorMap)

num_points= 1000
index = np.arange(num_points)
y = counts[:num_points]

fig, ax = plt.subplots(nrows=1, ncols=1)#prepareSubplot(np.arange(0, 6, 1), np.
 ↪arange(0, 14, 2))
plt.plot(index, counts[:1000], color=cmap(0), linewidth=3)

plt.xticks(ticks=np.arange(0, 1001, 100)) # set markers (ticks) and the␣
 ↪respective labels in the x-axis
plt.yticks(ticks=np.arange(0, 100001, 10000))
plt.xlabel('Paths')
plt.ylabel('Number of Hits')
plt.title('Visualizing paths')

plt.axhline(linewidth=2, color='#999999') # horizontal gray line (y=0)
plt.show()
```

```
/tmp/ipykernel_4389/72622161.py:14: MatplotlibDeprecationWarning: The get_cmap
function was deprecated in Matplotlib 3.7 and will be removed in 3.11. Use
```

```
``matplotlib.colormaps[name]`` or ``matplotlib.colormaps.get_cmap()`` or
``pyplot.get_cmap()`` instead.
  cmap = plt.cm.get_cmap(colorMap)
```



1.4.6 (3f) Example: Top Paths

For the final example, we'll find the top paths (URIs) in the log. Because we sorted `paths_df` for plotting, all we need to do is call `.show()` and pass in `n=10` and `truncate=False` as the parameters to show the top ten paths without truncating.

[40]:
```
# Top Paths
print ('Top Ten Paths:')
paths_df.show(n=10, truncate=False)

expected = [
  (u'/images/NASA-logosmall.gif', 97275),
  (u'/images/KSC-logosmall.gif', 75283),
  (u'/images/MOSAIC-logosmall.gif', 67356),
  (u'/images/USA-logosmall.gif', 66975),
  (u'/images/WORLD-logosmall.gif', 66351),
  (u'/images/ksclogo-medium.gif', 62670),
```

```
  (u'/ksc.html', 43618),
  (u'/history/apollo/images/apollo-logo1.gif', 37806),
  (u'/images/launch-logo.gif', 35119),
  (u'/', 30105)
]

testmti850.Test.assertEquals(paths_df.take(10), expected, 'incorrect Top Ten␣
 ↪Paths')
```

```
Top Ten Paths:
+----------------------------------------+-----+
|path                                    |count|
+----------------------------------------+-----+
|/images/NASA-logosmall.gif              |97275|
|/images/KSC-logosmall.gif               |75283|
|/images/MOSAIC-logosmall.gif            |67356|
|/images/USA-logosmall.gif               |66975|
|/images/WORLD-logosmall.gif             |66351|
|/images/ksclogo-medium.gif              |62670|
|/ksc.html                               |43618|
|/history/apollo/images/apollo-logo1.gif|37806|
|/images/launch-logo.gif                 |35119|
|/                                       |30105|
+----------------------------------------+-----+
only showing top 10 rows
1 test passed.
```

1.5 Part 4: Analyzing Web Server Log File

1.5.1 Now it is your turn to perform analyses on the web server log files.

(4a) Exercise: Top Ten Error Paths

What are the top ten paths which did not have return code 200? Create a sorted list containing the paths and the number of times that they were accessed with a non-200 return code and show the top ten.

Think about the steps that you need to perform to determine which paths did not have a 200 return code, how you will uniquely count those paths and sort the list.

[46]:
```
# TODO: Replace <FILL IN> with appropriate code

# You are welcome to structure your solution in a different way, so long as
# you ensure the variables used in the next Test section are defined

from pyspark.sql.functions import desc

# DataFrame containing all accesses that did not return a code 200
not200DF = logs_df.filter(logs_df["status"] != 200)
```

```
# # Sorted DataFrame containing all paths and the number of times they were␣
 ↪accessed with non-200 return code
logs_sum_df = not200DF.groupBy(not200DF["path"]).count().orderBy(desc("count"))

print ('Top Ten failed URLs:')
logs_sum_df.show(n=10, truncate=False)

print(logs_sum_df.count())
```

```
Top Ten failed URLs:
+----------------------------------------+-----+
|path                                    |count|
+----------------------------------------+-----+
|/images/NASA-logosmall.gif              |19072|
|/images/KSC-logosmall.gif               |11328|
|/images/MOSAIC-logosmall.gif            |8617 |
|/images/USA-logosmall.gif               |8565 |
|/images/WORLD-logosmall.gif             |8360 |
|/images/ksclogo-medium.gif              |7722 |
|/history/apollo/images/apollo-logo1.gif |4355 |
|/shuttle/countdown/images/countclock.gif|4227 |
|/images/launch-logo.gif                 |4178 |
|/                                       |3605 |
+----------------------------------------+-----+
only showing top 10 rows
9991
```

[47]:
```
# TEST Top ten error paths (4a)
top_10_err_urls = [(row[0], row[1]) for row in logs_sum_df.take(10)]
top_10_err_expected = [
  (u'/images/NASA-logosmall.gif', 19072),
  (u'/images/KSC-logosmall.gif', 11328),
  (u'/images/MOSAIC-logosmall.gif', 8617),
  (u'/images/USA-logosmall.gif', 8565),
  (u'/images/WORLD-logosmall.gif', 8360),
  (u'/images/ksclogo-medium.gif', 7722),
  (u'/history/apollo/images/apollo-logo1.gif', 4355),
  (u'/shuttle/countdown/images/countclock.gif', 4227),
  (u'/images/launch-logo.gif', 4178),
  (u'/', 3605)
]
testmti850.Test.assertEquals(logs_sum_df.count(), 9991, 'incorrect count for␣
 ↪logs_sum_df')
testmti850.Test.assertEquals(top_10_err_urls, top_10_err_expected, 'incorrect␣
 ↪Top Ten failed URLs')
```

```
1 test passed.
```

```
1 test passed.
```

1.5.2 (4b) Exercise: Number of Unique Hosts

How many unique hosts are there in the entire log?

There are multiple ways to find this. Try to find a more optimal way than grouping by 'host'.

[51]:
```
# TODO: Replace <FILL IN> with appropriate code

unique_host_count = logs_df.select("host").distinct().count()

print ('Unique hosts: {0}'.format(unique_host_count))
```

```
Unique hosts: 75060
```

[52]:
```
# TEST Number of unique hosts (4b)
testmti850.Test.assertEquals(unique_host_count, 75060, 'incorrect␣
 ↪unique_host_count')
```

```
1 test passed.
```

1.5.3 (4c) Exercise: Number of Unique Daily Hosts

For an advanced exercise, let's determine the number of unique hosts in the entire log on a day-by-day basis. This computation will give us counts of the number of unique daily hosts. We'd like a DataFrame sorted by increasing day of the month which includes the day of the month and the associated number of unique hosts for that day. Make sure you cache the resulting DataFrame `daily_hosts_df` so that we can reuse it in the next exercise.

Think about the steps that you need to perform to count the number of different hosts that make requests *each* day. *Since the log only covers a single month, you can ignore the month.* You may want to use the `dayofmonth` function in the `pyspark.sql.functions` module.

Description of each variable

`day_to_host_pair_df`

A DataFrame with two columns

column	explanation
`host`	the host name
`day`	the day of the month

There will be one row in this DataFrame for each row in `logs_df`. Essentially, you're just trimming and transforming each row of `logs_df`. For example, for this row in `logs_df`:

```
gw1.att.com - - [23/Aug/1995:00:03:53 -0400] "GET /shuttle/missions/sts-73/news HTTP/1.0" 302 -
```

your `day_to_host_pair_df` should have:

```
gw1.att.com 23
```

day_group_hosts_df

This DataFrame has the same columns as `day_to_host_pair_df`, but with duplicate (`day`, `host`) rows removed.

daily_hosts_df

A DataFrame with two columns:

column	explanation
`day`	the day of the month
`count`	the number of unique requesting hosts for that day

[125]:
```
# TODO: Replace <FILL IN> with appropriate code

from pyspark.sql.functions import dayofmonth

# logs_df.printSchema()

day_to_host_pair_df = logs_df.select(dayofmonth("time").alias("day"),"host")
# day_to_host_pair_df.show()

day_group_hosts_df = day_to_host_pair_df.dropDuplicates(["day", "host"])

daily_hosts_df = day_group_hosts_df.groupBy("day").count().orderBy("day")
daily_hosts_df.cache()

print ('Unique hosts per day:')

daily_hosts_df.show(n=30, truncate=False)
```

```
Unique hosts per day:

25/10/01 14:07:31 WARN CacheManager: Asked to cache already cached data.

+---+-----+
|day|count|
+---+-----+
|1  |2582 |
|3  |3222 |
|4  |4191 |
|5  |2502 |
|6  |2538 |
|7  |4108 |
|8  |4406 |
|9  |4317 |
|10 |4523 |
|11 |4346 |
|12 |2865 |
```

```
|13 |2650 |
|14 |4454 |
|15 |4214 |
|16 |4340 |
|17 |4385 |
|18 |4168 |
|19 |2550 |
|20 |2560 |
|21 |4135 |
|22 |4456 |
|23 |4368 |
|24 |4077 |
|25 |4407 |
|26 |2644 |
|27 |2690 |
|28 |4215 |
|29 |4826 |
|30 |5266 |
|31 |5916 |
+---+-----+
```

[126]:
```
# TEST Number of unique daily hosts (4c)
daily_hosts_list = (daily_hosts_df
                    .rdd.map(lambda r: (r[0], r[1]))
                    .take(30))
daily_hosts_list_expected = [(1, 2582), (3, 3222), (4, 4191), (5, 2502), (6,␣
 ↪2538), (7, 4108),
                             (8, 4406), (9, 4317), (10, 4523), (11, 4346), (12,␣
 ↪2865), (13, 2650),
                             (14, 4454), (15, 4214), (16, 4340), (17, 4385),␣
 ↪(18, 4168), (19, 2550),
                             (20, 2560), (21, 4135), (22, 4456), (23, 4368),␣
 ↪(24, 4077), (25, 4407),
                             (26, 2644), (27, 2690), (28, 4215), (29, 4826),␣
 ↪(30, 5266), (31, 5916)]
testmti850.Test.assertEquals(day_to_host_pair_df.count(), total_log_entries,␣
 ↪'incorrect row count for day_to_host_pair_df')
testmti850.Test.assertEquals(daily_hosts_df.count(), 30, 'incorrect␣
 ↪daily_hosts_df.count()')
testmti850.Test.assertEquals(daily_hosts_list, daily_hosts_list_expected,␣
 ↪'incorrect daily_hosts_df')
testmti850.Test.assertTrue(daily_hosts_df.is_cached, 'incorrect daily_hosts_df.
 ↪is_cached')
```

```
1 test passed.
```

```
1 test passed.
1 test passed.
1 test passed.
```

1.5.4 (4d) Exercise: Visualizing the Number of Unique Daily Hosts

Using the results from the previous exercise, we will use `matplotlib` to plot a line graph of the unique hosts requests by day. We need a list of days called `days_with_hosts` and a list of the number of unique hosts for each corresponding day called `hosts`.

WARNING: Simply calling `collect()` on your transformed DataFrame won't work, because `collect()` returns a list of Spark SQL `Row` objects. You must *extract* the appropriate column values from the `Row` objects. Hint: A loop will help.

[127]:
```
# TODO: Replace <FILL IN> with appropriate code

rows = daily_hosts_df.collect()
days_with_hosts = []
hosts = []

for row in rows:
    days_with_hosts.append(row[0])
    hosts.append(row[1])

print(days_with_hosts)
print(hosts)
```

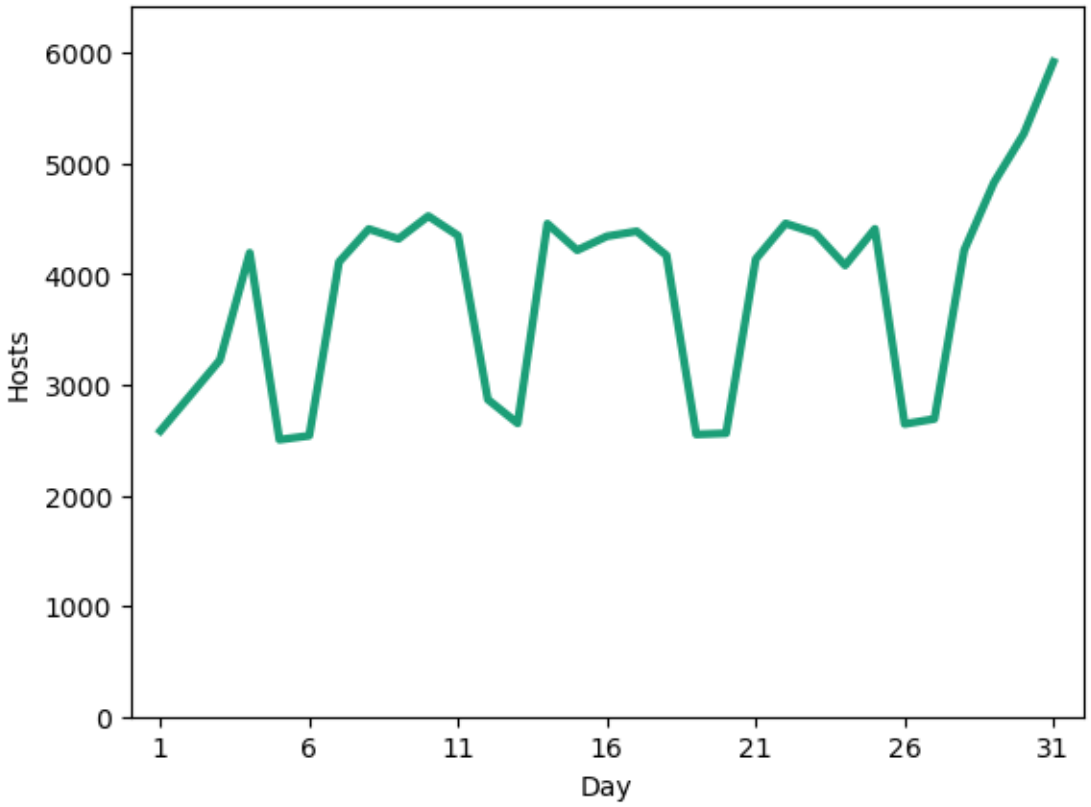
```
[1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23,
24, 25, 26, 27, 28, 29, 30, 31]
[2582, 3222, 4191, 2502, 2538, 4108, 4406, 4317, 4523, 4346, 2865, 2650, 4454,
4214, 4340, 4385, 4168, 2550, 2560, 4135, 4456, 4368, 4077, 4407, 2644, 2690,
4215, 4826, 5266, 5916]
```

[128]:
```
# TEST Visualizing unique daily hosts (4d)
# Possible problem here with Spark 4.0 [24/09/2025]
# To be updated
days_with_hosts_expected = list(range(1, 32))
days_with_hosts_expected.remove(2)

hosts_expected= [2582, 3222, 4191, 2502, 2538, 4108, 4406, 4317, 4523, 4346,␣
 ↪2865, 2650, 4454, 4214, 4340, 4385, 4168, 2550, 2560, 4135, 4456, 4368,␣
 ↪4077, 4407, 2644, 2690, 4215, 4826, 5266, 5916]

testmti850.Test.assertEquals(days_with_hosts, days_with_hosts_expected,␣
 ↪'incorrect days')
testmti850.Test.assertEquals(hosts, hosts_expected, 'incorrect hosts')
```

```
1 test passed.
1 test passed.
```

[65]:
```
# Possible problem here with Spark 4.0 [24/09/2025]
# To be updated
fig, ax = plt.subplots(nrows=1, ncols=1)

plt.xticks(ticks=np.arange(1, max(days_with_hosts) + 5, 5))
plt.yticks(ticks=np.arange(0, max(hosts) + 1000, 1000))
colorMap = 'Dark2'
cmap = plt.cm.get_cmap(colorMap)
plt.plot(days_with_hosts, hosts, color=cmap(0), linewidth=3)
plt.axis([0, max(days_with_hosts) + 1, 0, max(hosts) + 500])
plt.xlabel('Day')
plt.ylabel('Hosts')
plt.show()
```

```
/tmp/ipykernel_4389/3856202549.py:8: MatplotlibDeprecationWarning: The get_cmap
function was deprecated in Matplotlib 3.7 and will be removed in 3.11. Use
``matplotlib.colormaps[name]`` or ``matplotlib.colormaps.get_cmap()`` or
``pyplot.get_cmap()`` instead.
  cmap = plt.cm.get_cmap(colorMap)
```



1.5.5 (4e) Exercise: Average Number of Daily Requests per Host

Next, let's determine the average number of requests on a day-by-day basis. We'd like a list by increasing day of the month and the associated average number of requests per host for that day. Make sure you cache the resulting DataFrame `avg_daily_req_per_host_df` so that we can reuse it in the next exercise.

To compute the average number of requests per host, find the total number of requests per day (across all hosts) and divide that by the number of unique hosts per day (which we found in part 4c and cached as `daily_hosts_df`).

Since the log only covers a single month, you can skip checking for the month.

[105]:
```
# TODO: Replace <FILL IN> with appropriate code

total_req_per_day_df = logs_df.groupBy(dayofmonth("time").alias("day")).count()

# total_req_per_day_df.show()
# daily_hosts_df.show()


avg_daily_req_per_host_df = ( total_req_per_day_df
    .join(daily_hosts_df, "day")
    .withColumn("avg_reqs_per_host_per_day", total_req_per_day_df["count"] /␣
 ↪daily_hosts_df["unique_hosts"])
    .orderBy("day")
)

avg_daily_req_per_host_df = avg_daily_req_per_host_df.drop("count",␣
 ↪"unique_hosts")
avg_daily_req_per_host_df.cache()

print ('Average number of daily requests per Hosts is:\n')

avg_daily_req_per_host_df.show()
```

```
Average number of daily requests per Hosts is:

+---+------------------------+
|day|avg_reqs_per_host_per_day|
+---+------------------------+
|  1|       13.166537567776917|
|  3|       12.845437616387336|
|  4|       14.210689572894298|
|  5|       12.747002398081534|
|  6|        12.77383766745469|
|  7|         13.9634858812074|
|  8|       13.653427144802542|
|  9|        14.00463284688441|
```

```
| 10|       13.541454786646032|
| 11|       14.092498849516797|
| 12|       13.288307155322862|
| 13|       13.766037735849057|
| 14|       13.443646160754378|
| 15|       13.964641670621736|
| 16|         13.0536866359447|
| 17|       13.452223489167617|
| 18|       13.494721689059501|
| 19|       12.585882352941177|
| 20|             12.876171875|
| 21|         13.4316807738815|
+---+-------------------------+
only showing top 20 rows
```

[106]:
```
# TEST Average number of daily requests per hosts (4e)
from operator import itemgetter

avg_daily_req_per_host_list = (
  avg_daily_req_per_host_df.select('day',␣
 ↪avg_daily_req_per_host_df['avg_reqs_per_host_per_day'].cast('integer').
 ↪alias('avg_requests'))
                           .collect()
)

values = [(row[0], row[1]) for row in avg_daily_req_per_host_list]
print(values)

values_expected = [(1, 13), (3, 12), (4, 14), (5, 12), (6, 12), (7, 13), (8,␣
 ↪13), (9, 14), (10, 13), (11, 14), (12, 13), (13, 13), (14, 13), (15, 13),␣
 ↪(16, 13), (17, 13), (18, 13), (19, 12), (20, 12), (21, 13), (22, 12), (23,␣
 ↪13), (24, 12), (25, 13), (26, 11), (27, 12), (28, 13), (29, 14), (30, 15),␣
 ↪(31, 15)]
testmti850.Test.assertEquals(values, values_expected, 'incorrect␣
 ↪avgDailyReqPerHostDF')
testmti850.Test.assertTrue(avg_daily_req_per_host_df.is_cached, 'incorrect␣
 ↪avg_daily_req_per_host_df.is_cached')
```

```
[(1, 13), (3, 12), (4, 14), (5, 12), (6, 12), (7, 13), (8, 13), (9, 14), (10,
13), (11, 14), (12, 13), (13, 13), (14, 13), (15, 13), (16, 13), (17, 13), (18,
13), (19, 12), (20, 12), (21, 13), (22, 12), (23, 13), (24, 12), (25, 13), (26,
11), (27, 12), (28, 13), (29, 14), (30, 15), (31, 15)]
1 test passed.
1 test passed.
```

1.5.6 (4f) Exercise: Visualizing the Average Daily Requests per Unique Host

Using the result `avg_daily_req_per_host_df` from the previous exercise, use `matplotlib` to plot a line graph of the average daily requests per unique host by day.

`days_with_avg` should be a list of days and `avgs` should be a list of average daily requests (as integers) per unique hosts for each corresponding day. Hint: You will need to extract these from the Dataframe in a similar way to part 4d.

[129]:
```
# TODO: Replace <FILL IN> with appropriate code

rows = avg_daily_req_per_host_df.collect()

days_with_avg = []
avgs = []

for row in rows:
    days_with_avg.append(row[0])
    avgs.append(int(row[1]))

print(days_with_avg)

print(avgs)
```

```
[1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23,
24, 25, 26, 27, 28, 29, 30, 31]
[13, 12, 14, 12, 12, 13, 13, 14, 13, 14, 13, 13, 13, 13, 13, 13, 13, 12, 12, 13,
12, 13, 12, 13, 11, 12, 13, 14, 15, 15]
```

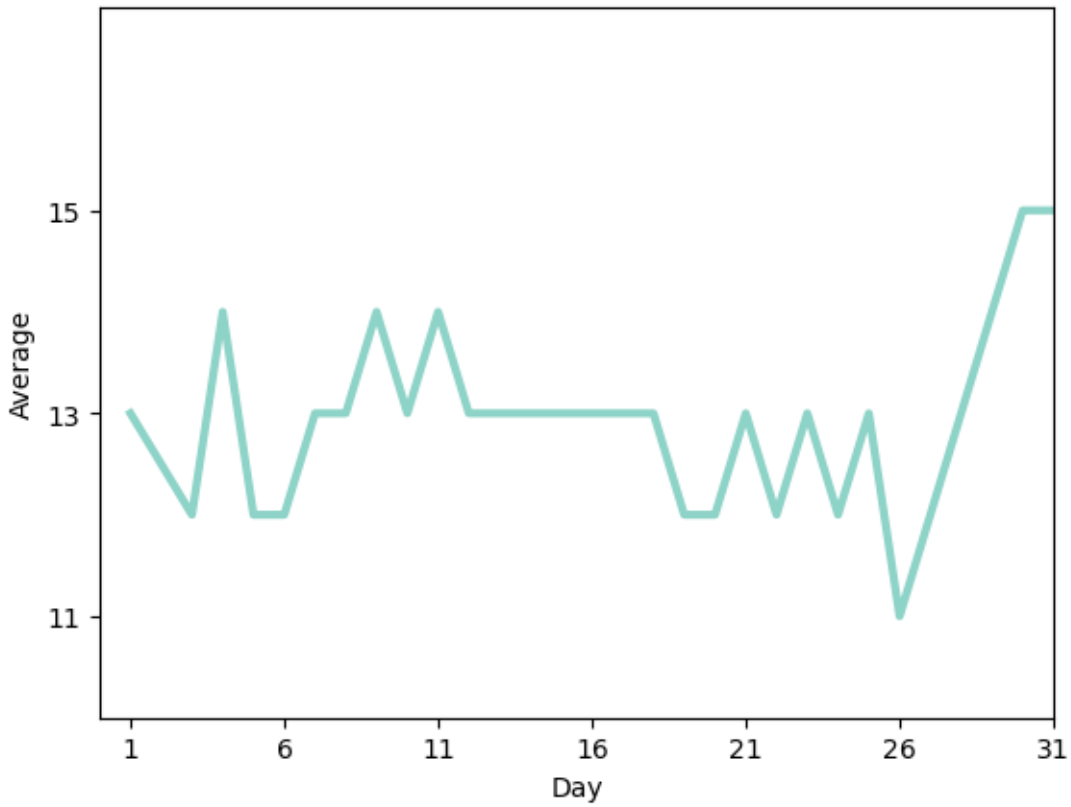
[130]:
```
# TEST Average Daily Requests per Unique Host (4f)
days_with_avg_expected = [1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16,␣
 ↪17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31]
avgs_expected = [13, 12, 14, 12, 12, 13, 13, 14, 13, 14, 13, 13, 13, 13, 13,␣
 ↪13, 13, 12, 12, 13, 12, 13, 12, 13, 11, 12, 13, 14, 15, 15]

testmti850.Test.assertEquals(days_with_avg, days_with_avg_expected, 'incorrect␣
 ↪days')
testmti850.Test.assertEquals([int(a) for a in avgs], avgs_expected, 'incorrect␣
 ↪avgs')
```

```
1 test passed.
1 test passed.
```

[131]:
```
fig, ax = plt.subplots(nrows=1, ncols=1)

plt.xticks(ticks=np.arange(1, max(days_with_avg) + 5, 5))
plt.yticks(ticks=np.arange(int(min(avgs)), max(avgs) + 2, 2))
colorMap = 'Set3'
cmap = plt.cm.get_cmap(colorMap)
```

```
plt.plot(days_with_avg, avgs, color=cmap(0), linewidth=3)
plt.axis([0, max(days_with_avg), 10, max(avgs) + 2])
plt.xlabel('Day')
plt.ylabel('Average')
plt.show()
```

```
/tmp/ipykernel_4389/3595012971.py:6: MatplotlibDeprecationWarning: The get_cmap
function was deprecated in Matplotlib 3.7 and will be removed in 3.11. Use
``matplotlib.colormaps[name]`` or ``matplotlib.colormaps.get_cmap()`` or
``pyplot.get_cmap()`` instead.
  cmap = plt.cm.get_cmap(colorMap)
```



1.6 Part 5: Exploring 404 Status Codes

Let's drill down and explore the error 404 status records. We've all seen those "404 Not Found" web pages. 404 errors are returned when the server cannot find the resource (page or object) the browser or client requested.

1.6.1 (5a) Exercise: Counting 404 Response Codes

Create a DataFrame containing only log records with a 404 status code. Make sure you `cache()` `not_found_df` as we will use it in the rest of this exercise.

How many 404 records are in the log?

[137]:
```
# TODO: Replace <FILL IN> with appropriate code
# logs_df.show()
not_found_df = logs_df.filter(logs_df["status"] == 404)
not_found_df.cache()

print( ('Found {0} 404 URLs').format(not_found_df.count()) )
```

```
Found 10056 404 URLs

25/10/01 14:12:13 WARN CacheManager: Asked to cache already cached data.
```

[138]:
```
# TEST Counting 404 (5a)
testmti850.Test.assertEquals(not_found_df.count(), 10056, 'incorrect␣
 ↪not_found_df.count()')
testmti850.Test.assertTrue(not_found_df.is_cached, 'incorrect not_found_df.
 ↪is_cached')
```

```
1 test passed.
1 test passed.
```

1.6.2 (5b) Exercise: Listing 404 Status Code Records

Using the DataFrame containing only log records with a 404 status code that you cached in (5a), print out a list up to 40 *distinct* paths that generate 404 errors.

No path should appear more than once in your list.

[140]:
```
# TODO: Replace <FILL IN> with appropriate code
not_found_paths_df = not_found_df.select("path")
# not_found_paths_df.show()

unique_not_found_paths_df = not_found_paths_df.dropDuplicates()

print ('404 URLS:\n')

unique_not_found_paths_df.show(n=40, truncate=False)
```

```
404 URLS:

+----------------------------------------------------------------+
|path                                                            |
+----------------------------------------------------------------+
|/shuttle/missions/sts-68/images/images.html                     |
|/history/apollo/a-001/news/                                     |
|/history/apollo/a-003/movies/                                   |
|/CSMT_PageNS                                                    |
|/missions/shuttle                                               |
|/shuttle/technology/sts-newsref/stsret-newsref/stsref-toc.html|
```

```
|/pub/wiinvn/win3/ww16_99_.zip                                    |
|/public.win3/winvn                                               |
|/shuttle/sts-1/sts-1-pa.jpg                                      |
|/history/apollo/apollo/13                                        |
|/shuttle/missions/sts-61-a/images/                               |
|/nssdc.gsfc.nasa.gov                                             |
|/shuttle/technology/images/sts-comm-small.gif                    |
|/shuttle/missions/sts-71/images/KSC-95EC-0916.txt                |
|/shuttle/countdown/ac.html                                       |
|/pub/winvn/docs                                                  |
|/IMAGES/RSS.GIF                                                  |
|/history/apollo/-apollo-13/apollo-3.html                         |
|/shuttle/missions/sts-71/./video/livevideo.jpeg                  |
|/thunder                                                         |
|/pub/winvn/readme.txt                                            |
|/ksc.shtml                                                       |
|/img/sportstalk3.gif                                             |
|/home.html                                                       |
|/shuttle/missions/sts-61a/mission-sts-61a.html                   |
|/shuttle/technology/sts-newsref/srb.html%23srb                   |
|/astronaut.*                                                     |
|/history/apollo-12/apollo-12.html                                |
|/elv/vidpicp.htlm                                                |
|/history/apollo/sa-9/images/                                     |
|/elv/FACILITIES/elvhead2.gif                                     |
|/shuttle/missions/sts-86/mission-sts-86.html                     |
|/history/gemini/gemini-12.html                                   |
|/histoty/apollo/aplool-13/apollo-13.html                         |
|/hqpao/hqpao-home.html                                           |
|/winvn/winvn.html.                                               |
|/www.quadralay.com                                               |
|/shuttle/missions/missionshtml                                   |
|/history/apollo/-apollo-13/apollo13.html                         |
|/shuttle/miccions/sts-73/mission-sts-73.html                     |
+-----------------------------------------------------------------+
only showing top 40 rows
```

[141]:
```
# TEST Listing 404 records (5b)

bad_unique_paths_40 = set([row[0] for row in unique_not_found_paths_df.
  ↪take(40)])
testmti850.Test.assertEquals(len(bad_unique_paths_40), 40, 'bad_unique_paths_40␣
  ↪not distinct')
```

```
1 test passed.
```

1.6.3 (5c) Exercise: Listing the Top Twenty 404 Response Code paths

Using the DataFrame containing only log records with a 404 response code that you cached in part (5a), print out a list of the top twenty paths that generate the most 404 errors.

Remember, top paths should be in sorted order

[158]:
```
# TODO: Replace <FILL IN> with appropriate code

# Pour cette cellule je pense qu'il y a un problème avec la fonction test -->␣
 ↪l'ordre attendu pour count = 77 ne suis pas l'ordre alphabétique

from pyspark.sql.functions import asc

top_20_not_found_df = not_found_paths_df.groupBy("path").count().
 ↪orderBy(desc("count"), asc("path")).limit(20)

print ('Top Twenty 404 URLs:\n')

top_20_not_found_df.show(n=20, truncate=False)
```

```
Top Twenty 404 URLs:

+-----------------------------------------------------------------------+-----+
|path                                                                   |count|
+-----------------------------------------------------------------------+-----+
|/pub/winvn/readme.txt                                                  |1337 |
|/pub/winvn/release.txt                                                 |1185 |
|/shuttle/missions/STS-69/mission-STS-69.html                           |682  |
|/images/nasa-logo.gif                                                  |319  |
|/shuttle/missions/sts-68/ksc-upclose.gif                               |251  |
|/elv/DELTA/uncons.htm                                                  |209  |
|/history/apollo/sa-1/sa-1-patch-small.gif                              |200  |
|/://spacelink.msfc.nasa.gov                                            |166  |
|/images/crawlerway-logo.gif                                            |160  |
|/history/apollo/a-001/a-001-patch-small.gif                            |154  |
|/history/apollo/pad-abort-test-1/pad-abort-test-1-patch-small.gif|144  |
|                                                                       |142  |
|/images/Nasa-logo.gif                                                  |85   |
|/history/apollo/images/little-joe.jpg                                  |84   |
|/shuttle/resources/orbiters/discovery.gif                              |82   |
|/shuttle/resources/orbiters/atlantis.gif                               |80   |
|/images/lf-logo.gif                                                    |77   |
|/robots.txt                                                            |77   |
|/shuttle/resources/orbiters/challenger.gif                             |77   |
|/pub                                                                   |57   |
+-----------------------------------------------------------------------+-----+
```

[156]:
```
# TEST Top twenty 404 URLs (5c)
top_20_not_found = [(row[0], row[1]) for row in top_20_not_found_df.take(20)]
top_20_expected = [
 (u'/pub/winvn/readme.txt', 1337),
 (u'/pub/winvn/release.txt', 1185),
 (u'/shuttle/missions/STS-69/mission-STS-69.html', 682),
 (u'/images/nasa-logo.gif', 319),
 (u'/shuttle/missions/sts-68/ksc-upclose.gif', 251),
 (u'/elv/DELTA/uncons.htm', 209),
 (u'/history/apollo/sa-1/sa-1-patch-small.gif', 200),
 (u'/://spacelink.msfc.nasa.gov', 166),
 (u'/images/crawlerway-logo.gif', 160),
 (u'/history/apollo/a-001/a-001-patch-small.gif', 154),
 (u'/history/apollo/pad-abort-test-1/pad-abort-test-1-patch-small.gif', 144),
 (u'', 142),
 (u'/images/Nasa-logo.gif', 85),
 (u'/history/apollo/images/little-joe.jpg', 84),
 (u'/shuttle/resources/orbiters/discovery.gif', 82),
 (u'/shuttle/resources/orbiters/atlantis.gif', 80),
 (u'/robots.txt', 77),
 (u'/images/lf-logo.gif', 77),
 (u'/shuttle/resources/orbiters/challenger.gif', 77),
 (u'/pub', 57)
]
testmti850.Test.assertEquals(top_20_not_found, top_20_expected, 'incorrect␣
 ↪top_20_not_found')
```

```
1 test failed. incorrect top_20_not_found
```

1.6.4 (5d) Exercise: Listing the Top Twenty-five 404 Response Code Hosts

Instead of looking at the paths that generated 404 errors, let's look at the hosts that encountered 404 errors. Using the DataFrame containing only log records with a 404 status codes that you cached in part (5a), print out a list of the top twenty-five hosts that generate the most 404 errors.

[164]:
```
# TODO: Replace <FILL IN> with appropriate code
# not_found_df.show()
hosts_404_count_df = not_found_df.select("host").groupBy("host").count().
 ↪orderBy(desc("count"))

print ('Top 25 hosts that generated errors:\n')

hosts_404_count_df.show(n=25, truncate=False)
```

```
Top 25 hosts that generated errors:

+--------------------------+-----+
|host                      |count|
+--------------------------+-----+
```

```
|dialip-217.den.mmc.com     |62   |
|piweba3y.prodigy.com       |47   |
|155.148.25.4               |44   |
|maz3.maz.net               |39   |
|gate.barr.com              |38   |
|m38-370-9.mit.edu          |37   |
|ts8-1.westwood.ts.ucla.edu |37   |
|nexus.mlckew.edu.au        |37   |
|204.62.245.32              |37   |
|scooter.pa-x.dec.com       |35   |
|reddragon.ksc.nasa.gov     |33   |
|www-c4.proxy.aol.com       |32   |
|piweba5y.prodigy.com       |31   |
|piweba4y.prodigy.com       |30   |
|www-d4.proxy.aol.com       |30   |
|internet-gw.watson.ibm.com |29   |
|unidata.com                |28   |
|163.206.104.34             |28   |
|spica.sci.isas.ac.jp       |27   |
|www-d2.proxy.aol.com       |26   |
|203.13.168.17              |25   |
|203.13.168.24              |25   |
|www-d1.proxy.aol.com       |23   |
|www-c2.proxy.aol.com       |23   |
|crl5.crl.com               |23   |
+---------------------------+-----+
only showing top 25 rows
```

[163]:
```
# TEST Top twenty-five 404 response code hosts (4d)

top_25_404 = [(row[0], row[1]) for row in hosts_404_count_df.take(25)]
top_25_404_expected = set([
    (u'dialip-217.den.mmc.com ', 62),
    (u'piweba3y.prodigy.com ', 47),
    (u'155.148.25.4 ', 44),
    (u'maz3.maz.net ', 39),
    (u'gate.barr.com ', 38),
    (u'204.62.245.32 ', 37),
    (u'nexus.mlckew.edu.au ', 37),
    (u'ts8-1.westwood.ts.ucla.edu ', 37),
    (u'm38-370-9.mit.edu ', 37),
    (u'scooter.pa-x.dec.com ', 35),
    (u'reddragon.ksc.nasa.gov ', 33),
    (u'www-c4.proxy.aol.com ', 32),
    (u'piweba5y.prodigy.com ', 31),
    (u'www-d4.proxy.aol.com ', 30),
    (u'piweba4y.prodigy.com ', 30),
```

```
    (u'internet-gw.watson.ibm.com ', 29),
    (u'unidata.com ', 28),
    (u'163.206.104.34 ', 28),
    (u'spica.sci.isas.ac.jp ', 27),
    (u'www-d2.proxy.aol.com ', 26),
    (u'203.13.168.17 ', 25),
    (u'203.13.168.24 ', 25),
    (u'www-c2.proxy.aol.com ', 23),
    (u'www-d1.proxy.aol.com ', 23),
    (u'crl5.crl.com ', 23)
])
testmti850.Test.assertEquals(len(top_25_404), 25, 'length of errHostsTop25 is␣
 ↪not 25')
testmti850.Test.assertEquals(len(set(top_25_404) - top_25_404_expected), 0,␣
 ↪'incorrect hosts_404_count_df')
```

```
1 test passed.
1 test passed.
```

1.6.5 (5e) Exercise: Listing 404 Errors per Day

Let's explore the 404 records temporally. Break down the 404 requests by day (cache the `errors_by_date_sorted_df` DataFrame) and get the daily counts sorted by day in `errors_by_date_sorted_df`.

Since the log only covers a single month, you can ignore the month in your checks.

[173]:
```
# TODO: Replace <FILL IN> with appropriate code
errors_by_date_sorted_df = not_found_df.groupBy(dayofmonth("time").
 ↪alias("day")).count().orderBy("day").cache()

print ('404 Errors by day:\n')

errors_by_date_sorted_df.show()
```

```
404 Errors by day:

+---+-----+
|day|count|
+---+-----+
|  1|  243|
|  3|  304|
|  4|  346|
|  5|  236|
|  6|  373|
|  7|  537|
|  8|  391|
|  9|  279|
| 10|  315|
```

```
| 11|  263|
| 12|  196|
| 13|  216|
| 14|  287|
| 15|  327|
| 16|  259|
| 17|  271|
| 18|  256|
| 19|  209|
| 20|  312|
| 21|  305|
+---+-----+
only showing top 20 rows
```

[174]:
```
# TEST 404 response codes per day (5e)

errors_by_date = [(row[0], row[1]) for row in errors_by_date_sorted_df.
 ↪collect()]
errors_by_date_expected = [
    (1, 243),
    (3, 304),
    (4, 346),
    (5, 236),
    (6, 373),
    (7, 537),
    (8, 391),
    (9, 279),
    (10, 315),
    (11, 263),
    (12, 196),
    (13, 216),
    (14, 287),
    (15, 327),
    (16, 259),
    (17, 271),
    (18, 256),
    (19, 209),
    (20, 312),
    (21, 305),
    (22, 288),
    (23, 345),
    (24, 420),
    (25, 415),
    (26, 366),
    (27, 370),
    (28, 410),
    (29, 420),
```

```
    (30, 571),
    (31, 526)
]
testmti850.Test.assertEquals(errors_by_date, errors_by_date_expected,␣
 ↪'incorrect errors_by_date_sorted_df')
testmti850.Test.assertTrue(errors_by_date_sorted_df.is_cached, 'incorrect␣
 ↪errors_by_date_sorted_df.is_cached')
```

```
1 test passed.
1 test passed.
```

1.6.6 (5f) Exercise: Visualizing the 404 Errors by Day

Using the results from the previous exercise, use `matplotlib` to plot a line or bar graph of the 404 response codes by day.

Hint: You'll need to use the same technique you used in (4f).

[175]:
```
# TODO: Replace <FILL IN> with appropriate code


rows = errors_by_date_sorted_df.collect()
days_with_errors_404 = []
errors_404_by_day = []

for row in rows:
    days_with_errors_404.append(row[0])
    errors_404_by_day.append(row[1])

print (days_with_errors_404)

print (errors_404_by_day)
```

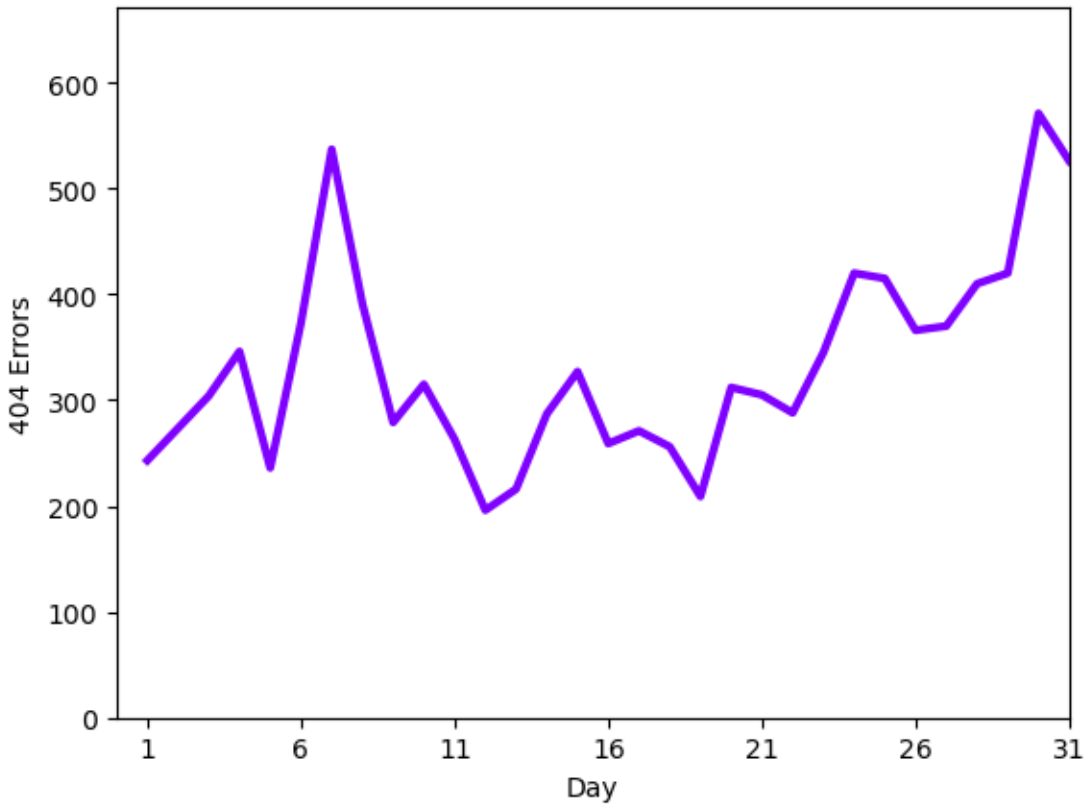
```
[1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23,
24, 25, 26, 27, 28, 29, 30, 31]
[243, 304, 346, 236, 373, 537, 391, 279, 315, 263, 196, 216, 287, 327, 259, 271,
256, 209, 312, 305, 288, 345, 420, 415, 366, 370, 410, 420, 571, 526]
```

[176]:
```
# TEST Visualizing the 404 Response Codes by Day (4f)
days_with_errors_404_expected = [1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14,␣
 ↪15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31]
errors_404_by_day_expected = [243, 304, 346, 236, 373, 537, 391, 279, 315, 263,␣
 ↪196, 216, 287, 327, 259, 271, 256, 209, 312, 305, 288, 345, 420, 415, 366,␣
 ↪370, 410, 420, 571, 526]
testmti850.Test.assertEquals(days_with_errors_404,␣
 ↪days_with_errors_404_expected, 'incorrect days_with_errors_404')
testmti850.Test.assertEquals(errors_404_by_day, errors_404_by_day_expected,␣
 ↪'incorrect errors_404_by_day')
```

```
1 test passed.
1 test passed.
```

[177]:
```
fig, ax = plt.subplots(nrows=1, ncols=1)
plt.xticks(ticks=np.arange(1, max(days_with_errors_404) + 5, 5))
plt.yticks(ticks=np.arange(0, max(errors_404_by_day) + 100, 100))
colorMap = 'rainbow'
cmap = plt.cm.get_cmap(colorMap)
plt.plot(days_with_errors_404, errors_404_by_day, color=cmap(0), linewidth=3)
plt.axis([0, max(days_with_errors_404), 0, max(errors_404_by_day) + 100])
plt.xlabel('Day')
plt.ylabel('404 Errors')
plt.show()
```

```
/tmp/ipykernel_4389/3859014693.py:5: MatplotlibDeprecationWarning: The get_cmap
function was deprecated in Matplotlib 3.7 and will be removed in 3.11. Use
``matplotlib.colormaps[name]`` or ``matplotlib.colormaps.get_cmap()`` or
``pyplot.get_cmap()`` instead.
  cmap = plt.cm.get_cmap(colorMap)
```



1.6.7 (5g) Exercise: Top Five Days for 404 Errors

Using the DataFrame `errors_by_date_sorted_df` you cached in the part (5e), what are the top five days for 404 errors and the corresponding counts of 404 errors?

[191]:
```
# TODO: Replace <FILL IN> with appropriate code
#errors_by_date_sorted_df.show(31)
top_err_date_df = errors_by_date_sorted_df.orderBy(desc("count")).limit(5)

print ('Top Five Dates for 404 Requests:\n')

top_err_date_df.show(5)
```

```
Top Five Dates for 404 Requests:

+---+-----+
|day|count|
+---+-----+
| 30|  571|
|  7|  537|
| 31|  526|
| 29|  420|
| 24|  420|
+---+-----+
```

[192]:
```
# TEST Five dates for 404 requests (4g)
top_err_date_list = [(col1, col2) for col1, col2 in top_err_date_df.take(5)]
top_err_date_list_expected = [(30, 571), (7, 537), (31, 526), (29, 420), (24,␣
 ↪420)]
testmti850.Test.assertEquals(top_err_date_list, top_err_date_list_expected,␣
 ↪'incorrect top_err_date_df')
```

```
1 test passed.
```

1.6.8 (5h) Exercise: Hourly 404 Errors

Using the DataFrame `not_found_df` you cached in the part (5a) and sorting by hour of the day in increasing order, create a DataFrame containing the number of requests that had a 404 return code for each hour of the day (midnight starts at 0). Cache the resulting DataFrame `hour_records_sorted_df` and print that as a list.

[201]:
```
# TODO: Replace <FILL IN> with appropriate code

from pyspark.sql.functions import hour

hour_records_sorted_df = not_found_df.groupBy(hour("time").alias("hour")).
 ↪count().orderBy("hour").cache()
```

```
print ('Top hours for 404 requests:\n')

hour_records_sorted_df.show(24)
```

```
Top hours for 404 requests:

+----+-----+
|hour|count|
+----+-----+
|   0|  344|
|   1|  327|
|   2|  600|
|   3|  363|
|   4|  183|
|   5|  160|
|   6|  135|
|   7|  218|
|   8|  340|
|   9|  359|
|  10|  492|
|  11|  428|
|  12|  651|
|  13|  614|
|  14|  522|
|  15|  549|
|  16|  550|
|  17|  586|
|  18|  425|
|  19|  440|
|  20|  432|
|  21|  434|
|  22|  430|
|  23|  474|
+----+-----+
```

[202]:
```
# TEST Hourly 404 response codes (5h)

errs_by_hour = [(col1, col2) for col1, col2 in hour_records_sorted_df.collect()]
for x in errs_by_hour:
    print(x)

errs_by_hour_expected = [
    (0, 344),
    (1, 327),
    (2, 600),
    (3, 363),
```

```
    (4, 183),
    (5, 160),
    (6, 135),
    (7, 218),
    (8, 340),
    (9, 359),
    (10, 492),
    (11, 428),
    (12, 651),
    (13, 614),
    (14, 522),
    (15, 549),
    (16, 550),
    (17, 586),
    (18, 425),
    (19, 440),
    (20, 432),
    (21, 434),
    (22, 430),
    (23, 474),
]
testmti850.Test.assertEquals(errs_by_hour, errs_by_hour_expected, 'incorrect␣
 ↪errs_by_hour')
testmti850.Test.assertTrue(hour_records_sorted_df.is_cached, 'incorrect␣
 ↪hour_records_sorted_df.is_cached')
```

```
(0, 344)
(1, 327)
(2, 600)
(3, 363)
(4, 183)
(5, 160)
(6, 135)
(7, 218)
(8, 340)
(9, 359)
(10, 492)
(11, 428)
(12, 651)
(13, 614)
(14, 522)
(15, 549)
(16, 550)
(17, 586)
(18, 425)
(19, 440)
(20, 432)
(21, 434)
```

```
(22, 430)
(23, 474)
1 test passed.
1 test passed.
```

1.6.9 (5i) Exercise: Visualizing the 404 Response Codes by Hour

Using the results from the previous exercise, use `matplotlib` to plot a line or bar graph of the 404 response codes by hour.

[203]:
```
# TODO: Replace <FILL IN> with appropriate code

rows = hour_records_sorted_df.collect()
hours_with_not_found = []
not_found_counts_per_hour = []

for row in rows :
    hours_with_not_found.append(row[0])
    not_found_counts_per_hour.append(row[1])

print(hours_with_not_found)

print(not_found_counts_per_hour)
```

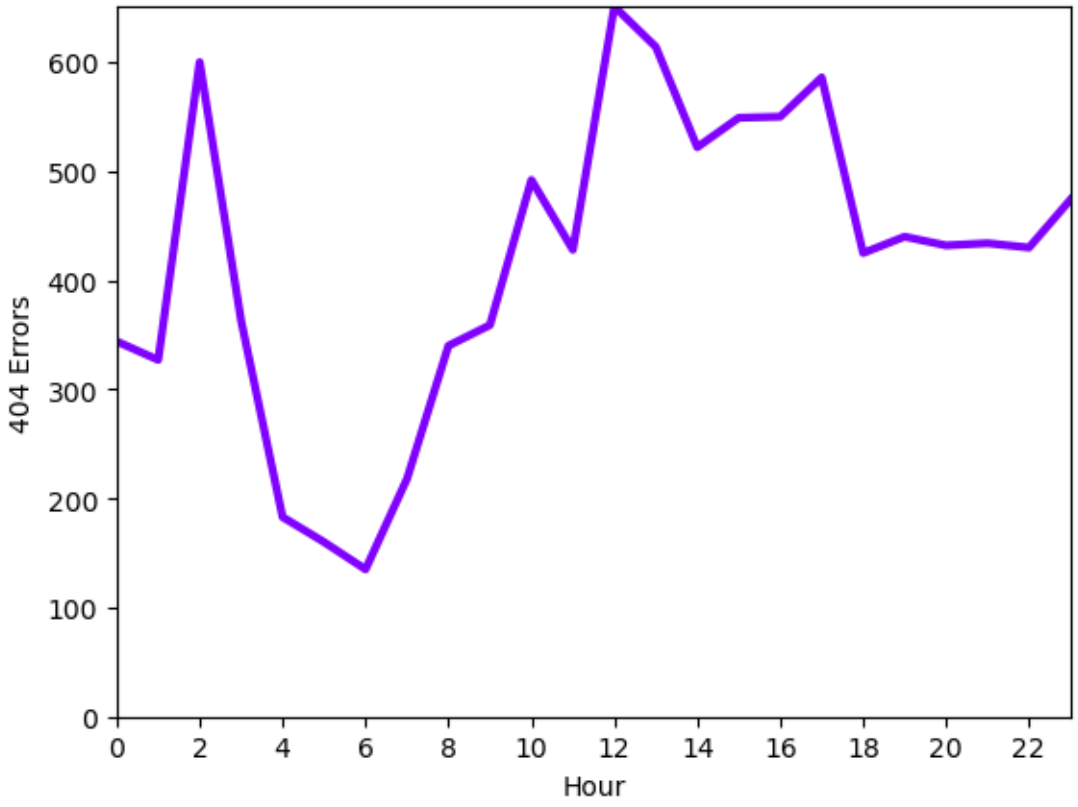
```
[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
22, 23]
[344, 327, 600, 363, 183, 160, 135, 218, 340, 359, 492, 428, 651, 614, 522, 549,
550, 586, 425, 440, 432, 434, 430, 474]
```

[204]:
```
# TEST Visualizing the 404 Response Codes by Hour (5i)
hours_with_not_found_expected = [0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13,␣
 ↪14, 15, 16, 17, 18, 19, 20, 21, 22, 23]
not_found_counts_per_hour_expected = [344, 327, 600, 363, 183, 160, 135, 218,␣
 ↪340, 359, 492, 428, 651, 614, 522, 549, 550, 586, 425, 440, 432, 434, 430,␣
 ↪474]
testmti850.Test.assertEquals(hours_with_not_found,␣
 ↪hours_with_not_found_expected, 'incorrect hours_with_not_found')
testmti850.Test.assertEquals(not_found_counts_per_hour,␣
 ↪not_found_counts_per_hour_expected, 'incorrect not_found_counts_per_hour')
```

```
1 test passed.
1 test passed.
```

[205]:
```
fig, ax = plt.subplots(nrows=1, ncols=1)
plt.xticks(ticks=np.arange(0, 24, 2))
plt.yticks(ticks=np.arange(0, max(not_found_counts_per_hour) + 100, 100))
colorMap = 'seismic'
plt.plot(hours_with_not_found, not_found_counts_per_hour, color=cmap(0),␣
 ↪linewidth=3)
```

```
plt.axis([0, max(hours_with_not_found), 0, max(not_found_counts_per_hour)])
plt.xlabel('Hour')
plt.ylabel('404 Errors')
plt.show()
```



1.7 Notebook Finished