John H Schulz

What is in all of those sstable files? Not just the data one, but all the rest too!

1	Bio and My Employer
2	Some background
3	The Test Table and data
4	The actual file contents
5	Conclusions and Other Stuff

Necessary Stuff

Nap time if you wish

BIO

Likes databases more than cheese sandwiches with peppered bacon and onions

Open source databases since 2003

Relational databases since 1984

Owned one of the first Apple IIs in the neighborhood

Lives in Michigan near the Detroit Zoo

About Pythian

11,400

Pythian currently manages more than 11,400 systems.

400+

Pythian currently employs more than 400 people in 200 cities in 35 countries

1997

Pythian was founded in 1997

Global Leader In IT Transformation And Operational Excellence

Unparalleled Expertise

- Top 5% in databases, applications, infrastructure, Big Data, Cloud, Data Science, and DevOps

Unmatched Certifications

- 9 Oracle ACEs, 4 Oracle ACE Directors, 1 Oracle ACE Associate
- 6 Microsoft MVPs, 1 Microsoft Certified Master
- 5 Google Platform Qualified Developers
- 1 Cloudera Champion of Big Data
- 1 Mongo DB Certified DBA Associate Level
- DataStax Certified Partner, 1 MVP, 3 Architect, 2 Certified administrators

Broad Technical Experience

- Oracle, Microsoft, MySQL, Oracle EBS, Hadoop, Cassandra, MongoDB, virtualization, configuration management, monitoring, trending, and more.

Background

Background

Some Terminology

Cassandra data directory structure

Some Terminology

Partition Key

The column(s) which are hashed to create a token mapped by the partitioner

Partition

The rows and columns which belong to a specific Partition Key

Clustering Key

The column(s) which, with the Partition Key define a row

Column

A specific entity with a type (can be user defined) a value, version, time to live and a name

More on columns

A single column can stand on its own in an SSTable file

What is the minimum information needed to find and use a column

- Column Name
- Data type
- Cluster column values
- Partition column values
- Version (usually a timestamp)
- Time to Live
- Value

The Cassandra data directory structure

data

- key-spaces
  - Tables
    - backups
    - snapshots
    - SSTable Files

Example partial directory tree from Cassandra 2.1

```
[root@Snoopy-3 node1]# ls -lR data
data:
total 12
drwxr-xr-x.  3 root root 4096 Aug 19 13:59 stuff
drwxr-xr-x. 19 root root 4096 Aug 19 13:58 system
drwxr-xr-x.  4 root root 4096 Aug 19 13:58 system_traces

data/stuff:
total 4
drwxr-xr-x. 2 root root 4096 Aug 19 14:00 simplefields-bdd61590663611e69c3e1d84c92693ab

data/stuff/simplefields-bdd61590663611e69c3e1d84c92693ab:
total 36
```

SSTable files in 2.1.9

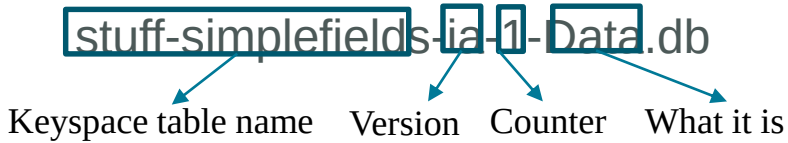
```
   8 Aug 19 14:00 stuff-simplefields-ka-1-CRC.db
1044 Aug 19 14:00 stuff-simplefields-ka-1-Data.db
   9 Aug 19 14:00 stuff-simplefields-ka-1-Digest.sha1
  24 Aug 19 14:00 stuff-simplefields-ka-1-Filter.db
 130 Aug 19 14:00 stuff-simplefields-ka-1-Index.db
4460 Aug 19 14:00 stuff-simplefields-ka-1-Statistics.db
 116 Aug 19 14:00 stuff-simplefields-ka-1-Summary.db
  79 Aug 19 14:00 stuff-simplefields-ka-1-TOC.txt
```

The SSTable files Cassandra 3.0

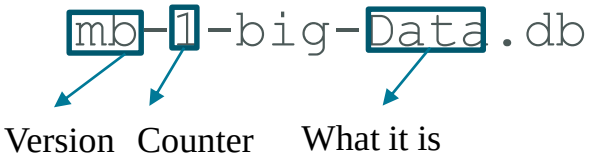
```
   8 Aug 19 10:42 mb-1-big-CRC.db
 481 Aug 19 10:42 mb-1-big-Data.db
  10 Aug 19 10:42 mb-1-big-Digest.crc32
  24 Aug 19 10:42 mb-1-big-Filter.db
  84 Aug 19 10:42 mb-1-big-Index.db
4799 Aug 19 10:42 mb-1-big-Statistics.db
  80 Aug 19 10:42 mb-1-big-Summary.db
  80 Aug 19 10:42 mb-1-big-TOC.txt
```

Parts of an SSTable file name

Pre 2.2



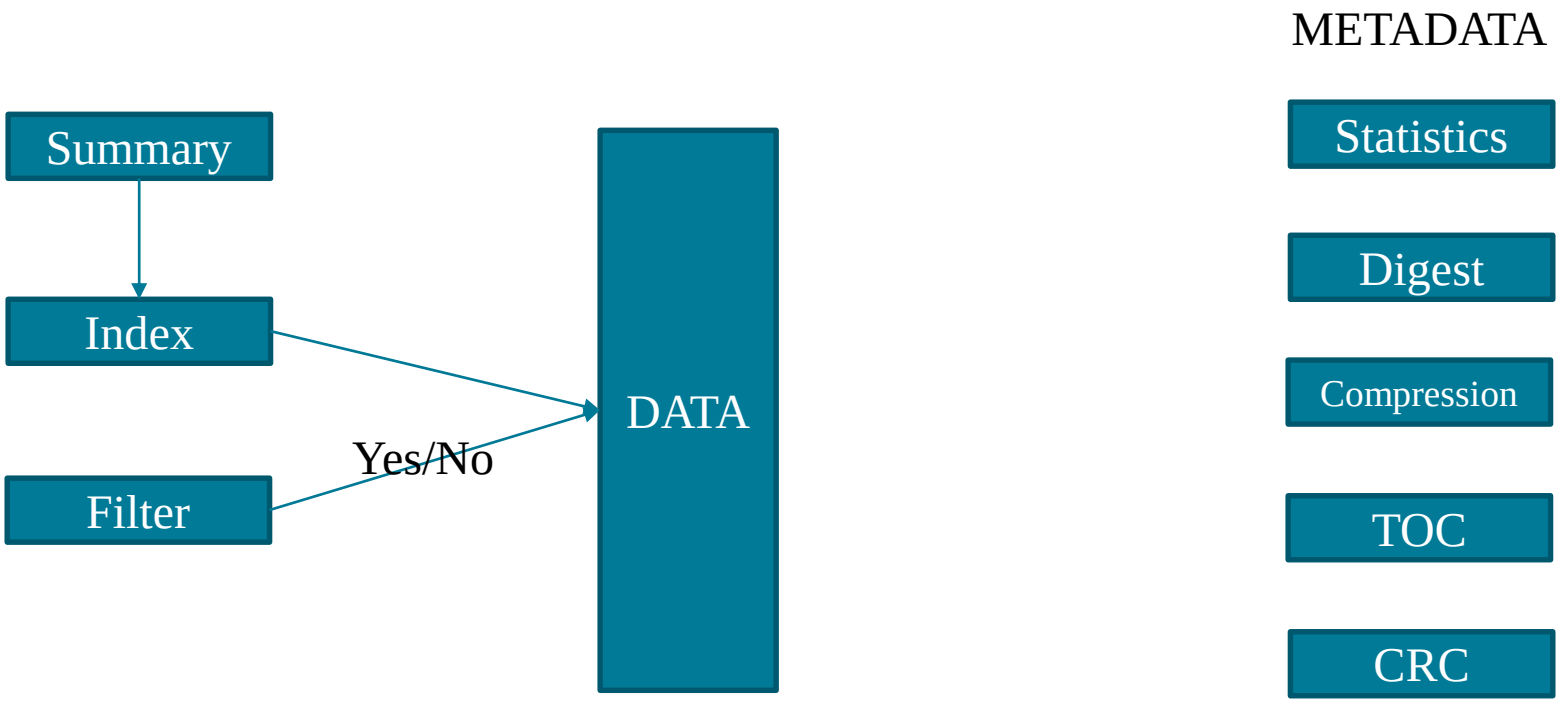
2.2 going forward



Data File Size Changed!!!

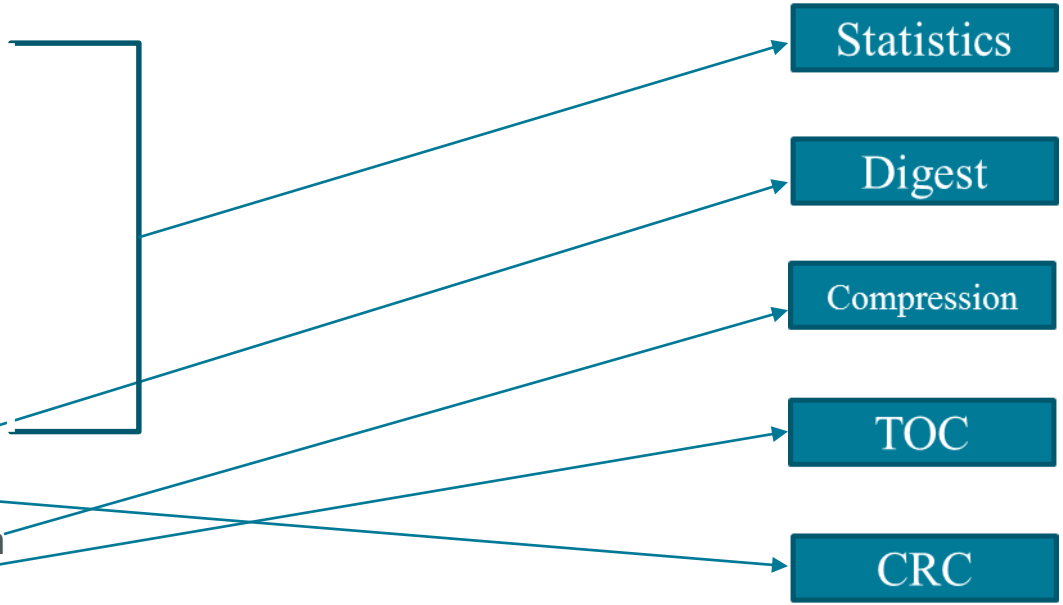
```
Cassandra 2.1
1044 Aug 19 14:00 stuff-simplefields-ka-1-Data.db


Cassandra 3.0
481 Aug 19 10:42 mb-1-big-Data.db
```

A relational view of how the files are connected



Metadata

- Partitioner
- Commit log location
- Field names and types
- Bloom filter chance
- Counter information
- Repair time
- Cardinality estimate
- Checksums
- Compression Algorithm
- Filenames



The Test Table and Data

My test table

```
CREATE TABLE stuff.simplefields (
    field1 text,
    field2 text,
    field3 text,
    field4 timestamp,
    field5 decimal,
    PRIMARY KEY ((field1, field2), field3)
) WITH CLUSTERING ORDER BY (field3 ASC)
AND compression = {};
```

Some test data to put in it

```
insert into simpleFields (field1, field2, field3, field4, field5)
         values('FV1', 'FV2' , 'FV3', '2016-08-11 11:37:25.976', '127.00');
insert into simpleFields (field1, field2, field3, field4, field5)
         values('GV1', 'GV2' , 'GV3', '2016-08-11 11:37:25.976', '127.10');
insert into simpleFields (field1, field2, field3, field4, field5)
         values('AV1', 'AV2' , 'AV3', '2016-08-11 11:37:25.976', '127.20');
insert into simpleFields (field1, field2, field3, field4, field5)
         values('BV1', 'BV2' , 'BV3', '2016-08-11 11:37:25.976', '127.30');
insert into simpleFields (field1, field2, field3, field4, field5)
         values('CV1', 'CV2' , 'CV3', '2016-08-11 11:37:25.976', '127.40');
insert into simpleFields (field1, field2, field3, field4, field5)
         values('CV1', 'CV2' , 'CV4', '2016-08-11 11:37:25.976', '127.50');
insert into simpleFields (field1, field2, field3, field4, field5)
         values('CV1', 'CV2' , 'CV5', '2016-08-11 11:37:25.976', '127.60');
insert into simpleFields (field1, field2, field3, field4, field5)
         values('CV1', 'CV2' , 'CV6', '2016-08-11 11:37:25.976', '127.70');
```

My Data in the Table

field1	field2	field3	field4	field5
CV1	CV2	CV3	2016-08-11 11:37:25.976	127.40
CV1	CV2	CV4	2016-08-11 11:37:25.976	127.50
CV1	CV2	CV5	2016-08-11 11:37:25.976	127.60
CV1	CV2	CV6	2016-08-11 11:37:25.976	127.70
FV1	FV2	FV3	2016-08-11 11:37:25.976	127.00
GV1	GV2	GV3	2016-08-11 11:37:25.976	127.10
BV1	BV2	BV3	2016-08-11 11:37:25.976	127.30
AV1	AV2	AV3	2016-08-11 11:37:25.976	127.20

How I got my clusters going

Using CCM – Cassandra Cluster Manager

https://github.com/pcmanus/ccm

Example of creating a cluster:

```
ccm create --vnodes --nodes=1 –v 2.1.9--start ver-219
ccm node1 nodetool status
Datacenter: datacenter1
=======================
Status=Up/Down
|/ State=Normal/Leaving/Joining/Moving
--  Address    Load       Tokens  Owns    Host ID                               Rack
UN  127.0.0.1  196.38 KB  256     ?       6ecec1cd-424d-4173-a83f-aefe485f4e2c  rack1
```

Clusters currently on my Machine

ver-12a	ver-219	ver227	ver_33
ver-12a-multi	ver-219-single	ver-227-single	
ver_20_11a	ver-221	ver-301	
test	ver-2011-single	ver_225	ver-308
ver-120-single	ver-213-multi	ver_225_6	ver-308-single

A peek inside the files

Quick overview of each File

CRC – A single CRC of the Data file

Compression – The name of the compression engine used

Data – The SSTable data stored in sorted order by Partition/Cluster Key

Digest – One or more CRCs one per 2GB segment of the Data file

Filter – The bloom filter – folded hashes of all key values in the Data file

Index – A list of all partition/Cluster Keys

Statistics – Hodgepodge of metadata about the SSTable

Summary – A list of some of the Partition/Cluster keys – Think second level index

TOC – Text list of the files

A peek inside the less interesting files

hexdump mb-1-big-CRC.db

0000000 0100 0000 c7ff 8b91

0000008

cat mb-1-big-Digest.crc32

4291269003

cat mb-1-big-TOC.txt

Filter.db

CRC.db

Data.db

Digest.crc32

Summary.db

TOC.txt

Statistics.db

Index.db

Cassandra 2.1 Index

```
0000000  \0  \f  \0 003   C   V   1 \0  \0 003   C   V   2 \0  \0  \0
0000010  \0  \0  \0  \0  \0  \0  \0  \0  \0  \0 \0  \f  \0 003   F   V
0000020   1 \0  \0 003   F   V   2 \0  \0  \0  \0  \0  \0  \0 001 340
0000030  \0  \0  \0  \0  \0  \f  \0 003   G   V   1  \0  \0 003   G   V
0000040   2  \0  \0  \0  \0  \0  \0  \0 002   m  \0  \0  \0  \0  \0  \f
0000050  \0 003   B   V   1  \0  \0 003   B   V   2  \0  \0  \0  \0  \0
0000060  \0  \0 002 372  \0  \0  \0  \0  \0  \f \0 003   A   V   1 \0
0000070  \0 003   A   V   2 \0  \0  \0  \0  \0  \0  \0 003 207  \0  \0
0000080  \0  \0
```

Cassandra 3.0 Index

```
hexdump -c mb-1-big-Index.db
0000000  \0  \f  \0 003   C   V   1  \0  \0 003   C   V   2  \0  \0  \0
0000010  \0  \f  \0 003   F   V   1  \0  \0 003   F   V   2  \0 200 312
0000020  \0  \0  \f  \0 003   G   V   1  \0  \0 003   G   V   2  \0 201
0000030 017  \0  \0  \f  \0 003   B   V   1  \0  \0 003   B   V   2  \0
0000040 201   U  \0  \0  \f  \0 003   A   V   1  \0  \0 003   A   V   2
0000050  \0 201 233  \0
```

Cassandra 2.1 Summary

```
0000000  \0  \0  \0 200  \0  \0  \0 001  \0  \0  \0  \0  \0  \0  \0 030
0000010  \0  \0  \0 200  \0  \0  \0 001 004  \0  \0  \0 \0 003    C   V
0000020   1 \0 \0 003    C   V   2 \0  \0  \0  \0  \0  \0  \0  \0  \0
0000030  \0  \0  \0  \f  \0 003   C   V   1  \0  \0 003   C   V   2  \0
0000040  \0  \0  \0  \f  \0 003   A   V   1  \0  \0 003   A   V   2  \0
0000050  \0 004   m   m   a   p  \0  \0  \0 001  \0  \0  \0  \0  \0  \0
0000060  \0  \0  \0 004   m   m   a   p  \0  \0  \0 001  \0  \0  \0  \0
0000070  \0  \0  \0  \0
```

CASSANDRA SUMMIT 2016

Cassandra 3.0 Summary

```
hexdump -c mb-1-big-Summary.db
0000000  \0  \0  \0 200  \0  \0  \0 001  \0  \0  \0  \0  \0  \0  \0 030
0000010  \0  \0  \0 200  \0  \0  \0 001 004  \0  \0  \0 \0 003   C   V
0000020   1 \0 \0 003   C   V   2 \0  \0  \0  \0  \0  \0  \0  \0  \0
0000030  \0  \0  \0  \f  \0 003   C   V   1  \0 \0 003   C   V   2  \0
0000040  \0  \0  \0  \f  \0 003   A   V   1 \0  \0 003   A   V   2 \0
```

Cassandra 3.0 Statistics Part Beginning

```
hexdump -c mb-1-big-Statistics.db
0000000  \0  \0  \0 004  \0  \0  \0  \0  \0  \0  \0    $  \0  \0  \0 001
0000010  \0  \0  \0    Y  \0  \0  \0 002  \0  \0  \0    y  \0  \0  \0 003
0000020  \0  \0 021 250  \0    +    o    r    g    .    a    p    a    c    h    e
0000030    .    c    a    s    s    a    n    d    r    a    .    d    h    t    .    M
0000040    u    r    m    u    r    3    P    a    r    t    i    t    i    o    n    e
0000050    r    ?  204    z  341    G  256  024    {  \0  \0  \0  034  377  377  377
0000060  376   \r  031  001  005  256  346  275  001  302  263  326  005  334  361  364
0000070   \t  354  241  224   \n  324  234  337  001  \0  \0  \0  227  \0  \0  \0
0000080  \0  \0  \0  \0  001  \0  \0  \0  \0  \0  \0  \0  \0  \0  \0  \0
```

Cassandra 3.0 Statistics Tail end

```
00001230  70 65 29 01 28 6f 72 67  2e 61 70 61 63 68 65 2e  |pe).(org.apache.|
00001240  63 61 73 73 61 6e 64 72  61 2e 64 62 2e 6d 61 72  |cassandra.db.mar|
00001250  73 68 61 6c 2e 55 54 46  38 54 79 70 65 00 02 06  |shal.UTF8Type...|
00001260  66 69 65 6c 64 34 28 6f  72 67 2e 61 70 61 63 68  |field4(org.apach|
00001270  65 2e 63 61 73 73 61 6e  64 72 61 2e 64 62 2e 6d  |e.cassandra.db.m|
00001280  61 72 73 68 61 6c 2e 55  54 46 38 54 79 70 65 06  |arshal.UTF8Type.|
00001290  66 69 65 6c 64 35 28 6f  72 67 2e 61 70 61 63 68  |field5(org.apach|
000012a0  65 2e 63 61 73 73 61 6e  64 72 61 2e 64 62 2e 6d  |e.cassandra.db.m|
000012b0  61 72 73 68 61 6c 2e 55  54 46 38 54 79 70 65     |arshal.UTF8Type|
```

Cassandra 2.1 First Part

```
0000000  \0  \f  \0 003   C   V   1  \0  \0 003   C   V   2  \0 177 377
0000010 377 377 200  \0  \0  \0  \0  \0  \0  \0  \0  \t  \0 003   C   V
0000020    3  \0  \0  \0  \0  \0  \0  005    :   p   w   w   [ 341  \0  \0
0000030  \0  \0  \0 017  \0 003   C   V   3  \0  \0 006   f   i   e   l
0000040    d   4  \0  \0  \0 005    :   p   w   w   [ 341  \0  \0  \0 027
0000050    2   0   1   6   -   0   8   -   1   1       1   1   :   3   7
0000060    :   2   5   .   9   7   6  \0 017  \0 003   C   V   3  \0  \0
0000070 006   f   i   e   l   d   5  \0  \0  \0 005   :   p   w   w   [
0000080 341  \0  \0  \0 006   1   2   7   .   4   0  \0  \t  \0 003   C
0000090   V   4  \0  \0  \0  \0  \0  \0 005   :   p   w   w   n   /  \0
00000a0  \0  \0  \0  \0 017  \0 003   C   V   4  \0  \0 006   f   i   e
00000b0   l   d   4  \0  \0  \0 005   :   p   w   w   n   /  \0  \0  \0
00000c0 027   2   0   1   6   -   0   8   -   1   1       1   1   :   3
00000d0   7   :   2   5   .   9   7   6  \0 017  \0 003   C   V   4  \0
00000e0  \0 006   f   i   e   l   d   5  \0  \0  \0 005   :   p   w   w
00000f0   n   /  \0  \0  \0 006   1   2   7   .   5   0  \0  \t  \0 003
0000100   C   V   5  \0  \0  \0  \0  \0  \0 005   :   p   w   w 200 261
0000110  \0  \0  \0  \0  \0 017  \0 003   C   V   5  \0  \0 006   f   i
0000120   e   l   d   4  \0  \0  \0 005   :   p   w   w 200 261  \0  \0
0000130  \0 027   2   0   1   6   -   0   8   -   1   1       1   1   :
```

Cassandra 2.1 Second Part

```
0000140   3   7   :   2   5   .   9   7   6  \0 017  \0 003   C   V   5
0000150  \0  \0 006   f   i   e   l   d   5  \0  \0  \0 005   :   p   w
0000160   w 200 261  \0  \0  \0 006   1   2   7   .   6   0  \0  \t  \0
0000170 003   C   V   6  \0  \0  \0  \0  \0  \0 005   :   p   w   w 213
0000180  \a  \0  \0  \0  \0  \0 017  \0 003   C   V   6  \0  \0 006   f
0000190   i   e   l   d   4  \0  \0  \0 005   :   p   w   w 213  \a  \0
00001a0  \0  \0 027   2   0   1   6   -   0   8   -   1   1       1   1
00001b0   :   3   7   :   2   5   .   9   7   6  \0 017  \0 003   C   V
00001c0   6  \0  \0 006   f   i   e   l   d   5  \0  \0  \0 005   :   p
00001d0   w   w 213  \a  \0  \0  \0 006   1   2   7   .   7   0  \0  \0
00001e0  \0  \f  \0 003   F   V   1  \0  \0 003   F   V   2  \0 177 377
00001f0 377 377 200  \0  \0  \0  \0  \0  \0  \0  \0  \t  \0 003   F   V
0000200   3  \0  \0  \0  \0  \0  \0 005   :   p   w   v 271   d  \0  \0
0000210  \0  \0  \0 017  \0 003   F   V   3  \0  \0 006   f   i   e   l
0000220   d   4  \0  \0  \0 005   :   p   w   v 271   d  \0  \0  \0 027
0000230   2   0   1   6   -   0   8   -   1   1       1   1   :   3   7
0000240   :   2   5   .   9   7   6  \0 017  \0 003   F   V   3  \0  \0
0000250 006   f   i   e   l   d   5  \0  \0  \0 005   :   p   w   v 271
0000260   d  \0  \0  \0 006   1   2   7   .   0   0  \0  \0  \0  \f  \0
```

Cassandra 2.1 Third Part

```
0000270 003   G   V   1  \0  \0 003   G   V   2  \0 177 377 377 377 200
0000280  \0  \0  \0  \0  \0  \0  \0  \0  \t  \0 003   G   V   3  \0  \0
0000290  \0  \0  \0  \0 005   :   p   w   w   : 255  \0  \0  \0  \0  \0
00002a0 017  \0 003   G   V   3  \0  \0 006   f   i   e   l   d   4  \0
00002b0  \0  \0 005   :   p   w   w   : 255  \0  \0  \0 027   2   0   1
00002c0   6   -   0   8   -   1   1       1   1   :   3   7   :   2   5
00002d0   .   9   7   6  \0 017  \0 003   G   V   3  \0  \0 006   f   i
00002e0   e   l   d   5  \0  \0  \0 005   :   p   w   w   : 255  \0  \0
00002f0  \0 006   1   2   7   .   1   0  \0  \0  \0  \f  \0 003   B   V
0000300   1  \0  \0 003   B   V   2  \0 177 377 377 377 200  \0  \0  \0
0000310  \0  \0  \0  \0  \0  \t  \0 003   B   V   3  \0  \0  \0  \0  \0
0000320  \0 005   :   p   w   w   Q 234  \0  \0  \0  \0  \0 017  \0 003
0000330   B   V   3  \0  \0 006   f   i   e   l   d   4  \0  \0  \0 005
0000340   :   p   w   w   Q 234  \0  \0  \0 027   2   0   1   6   -   0
0000350   8   -   1   1       1   1   :   3   7   :   2   5   .   9   7
0000360   6  \0 017  \0 003   B   V   3  \0  \0 006   f   i   e   l   d
0000370   5  \0  \0  \0 005   :   p   w   w   Q 234  \0  \0  \0 006   1
0000380   2   7   .   3   0  \0  \0  \0  \f  \0 003   A   V   1  \0  \0
0000390 003   A   V   2  \0 177 377 377 377 200  \0  \0  \0  \0  \0  \0
```

Cassandra 3.0 Part A

```
0000000  \0  \f  \0 003   C   V   1  \0  \0 003   C   V   2  \0 177 377
0000010 377 377 200  \0  \0  \0  \0  \0  \0   $  \0 003   C   V   3
0000020   $ 032 270 221  \b 027   2   0   1   6   -   0   8   -   1   1
0000030       1   1   :   3   7   :   2   5   .   9   7   6  \b 006   1
0000040   2   7   .   4   0   $  \0 003   C   V   4   %   + 300   D  \a
0000050  \b 027   2   0   1   6   -   0   8   -   1   1       1   1   :
0000060   3   7   :   2   5   .   9   7   6  \b 006   1   2   7   .   5
0000070   0   $  \0 003   C   V   5   %   , 300   N 234  \b 027   2   0
0000080   1   6   -   0   8   -   1   1       1   1   :   3   7   :   2
0000090   5   .   9   7   6  \b 006   1   2   7   .   6   0   $  \0 003
00000a0   C   V   6   %   , 300   [   `  \b 027   2   0   1   6   -   0
00000b0   8   -   1   1       1   1   :   3   7   :   2   5   .   9   7
00000c0   6  \b 006   1   2   7   .   7   0 001  \0  \f  \0 003   F   V
00000d0   1  \0  \0 003   F   V   2  \0 177 377 377 377 200  \0  \0  \0
00000e0  \0  \0  \0  \0   $  \0 003   F   V   3   # 032  \0  \b 027   2
00000f0   0   1   6   -   0   8   -   1   1       1   1   :   3   7   :
0000100   2   5   .   9   7   6  \b 006   1   2   7   .   0   0 001  \0
0000110  \f  \0 003   G   V   1  \0  \0 003   G   V   2  \0 177 377 377
```

Cassandra 3.0 Part B

```
0000120 377 200  \0  \0  \0  \0  \0  \0  \0   $  \0 003   G   V   3   $
0000130 032 221 346  \b 027   2   0   1   6   -   0   8   -   1   1
0000140   1   1   :   3   7   :   2   5   .   9   7   6  \b 006   1   2
0000150   7   .   1   0 001  \0  \f  \0 003   B   V   1  \0  \0 003   B
0000160   V   2  \0 177 377 377 377 200  \0  \0  \0  \0  \0  \0  \0   $
0000170  \0 003   B   V   3   $ 032 255 036  \b 027   2   0   1   6   -
0000180   0   8   -   1   1       1   1   :   3   7   :   2   5   .   9
0000190   7   6  \b 006   1   2   7   .   3   0 001  \0  \f  \0 003   A
00001a0   V   1  \0  \0 003   A   V   2  \0 177 377 377 377 200  \0  \0
00001b0  \0  \0  \0  \0  \0   $  \0 003   A   V   3   $ 032 240   >  \b
00001c0 027   2   0   1   6   -   0   8   -   1   1       1   1   :   3
00001d0   7   :   2   5   .   9   7   6  \b 006   1   2   7   .   2   0
00001e0 001
```

Bloom Filter contents 2.1 and 3.0

```
hexdump stuff-simplefields-ka-1-Filter.db
0000000 0000 0500 0000 0200 0032 2090 3002 0009
0000010 1803 0110 0022 40c0

hexdump mb-1-big-Filter.db
0000000 0000 0500 0000 0200 1808 5068 8822 0019
0000010 1012 1100 0002 4180
```

Conclusions References and questions

Conclusions

- Cassandra 3.0 SSTable format is considerably more compact
  - In Cassandra 3.0 field names are no longer stored in the data file field Ids are used instead
  - TTL and Timestamps are grouped
  - TTL and Timestamp Differences are stored as deltas
- Sorted String Tables are sorted…. By Partition Key Token Value
- The Statistics file is full of a bunch of different stuff
- Data, Summary, Filter and Statistics have important useful stuff

References

http://distributeddatastore.blogspot.com/2013/08/cassandra-sstable-storage-format.html

https://github.com/scylladb/scylla/wiki/Storage-On-Disk-File-Formats

http://thelastpickle.com/blog/2016/03/04/introductiont-to-the-apache-cassandra-3-storage-engine.html

QUESTIONS?

In version "ka", contents of Statistics component file is split in to 3 types of metadata called validation, compaction and stats. Validation metadata is used to validate SSTable which includes partitioner and bloom filter fp chance fields. Compaction metadata includes ancestors information which is also available in older formats and a new field called cardinality estimator. Cardinality estimator is used to efficiently pre-allocate bloom filter space in a merged compaction file by estimating how much the input SStables overlap. Stats metadata contains rest of the information available in old formats and two additional fields. First one is a flag to track the presence of local/remote counter shards and the other one is for storing the repair time.