FAST5文件是牛津纳米孔技术（Oxford Nanopore Technologies, ONT）测序平台特有的一种文件格式，用于存储测序过程中产生的原始电信号数据。这种格式的文件包含了测序过程中的全部输出信息，如捕获的电信号值、设备运行时间、电压、温度等。FAST5是HDF5文件格式的一个变种，HDF5（Hierarchical Data Format version 5）是一种设计用于存储和组织大量数据的文件格式，通常具有.hdf5或.h5扩展名。

FAST5文件的数据结构复杂，可以包含许多内容，并且可以继续添加。它类似于一个经过压缩的文件夹，里面包含许多文件。每个纳米孔测序完成后，可以输出一个FAST5文件。一个MinION的flowcell可以生成10-20Gb的数据量，包含几百万条reads。FAST5文件的优点是内容非常全面，可以存储所有的信息，但缺点是占用空间特别大。例如，23M左右的碱基序列，存储为FASTQ格式大约需要45M空间，而原始的FAST5文件则可能需要613M的存储空间，大约是FASTQ的30倍。

FAST5文件可以通过软件如HDFView进行查看，它提供了一个用户友好的界面来浏览和查看HDF5文件的内容。此外，可以使用Python中的h5py库来处理FAST5文件，这允许用户编程方式访问和操作文件中的数据。

在处理FAST5文件时，有时可能需要将多个FAST5文件合并为一个文件，或者将一个文件拆分为多个文件。这可以通过使用single\_to\_multi\_fast5和multi\_to\_single\_fast5工具来完成。

在FAST5文件的上下文中，"schemas"通常指的是用于描述和验证文件内容的数据模型或结构定义。这些模式定义了文件中可以包含的数据类型、属性、关系以及它们应该如何组织。在HDF5和FAST5文件中，schemas可能以特定的格式存储在文件的元数据中，用于确保数据的一致性和可读性。

在FAST5文件中，数据被组织成多个子字段（subfields），这些子字段包含了测序过程中产生的各种信息。FAST5文件是HDF5格式的，因此它们可以包含复杂的数据结构，如组（groups）、数据集（datasets）、属性（attributes）等。以下是FAST5文件中一些常见的数据子字段：

1. \*\*Raw Data\*\*:

- `Raw/Reads`: 包含原始的电信号读数，这是测序过程中捕获的电信号的原始数据。

2. \*\*Signal Processing\*\*:

- `Signal/processing`: 包含信号处理的参数和结果，如基线漂移校正、噪声滤波等。

3. \*\*Alignment Information\*\*:

- `Alignment`: 如果进行了序列比对，这里会包含比对的相关信息。

4. \*\*Event Detection\*\*:

- `Events`: 包含事件检测的结果，如模板跳跃、错误校正等。

5. \*\*Basecalling\*\*:

- `Basecall\_1D`: 包含一维基频分析的结果，这是将电信号转换为碱基呼叫的第一步。

- `Basecall\_2D`: 包含二维基频分析的结果，用于提高碱基呼叫的准确性。

6. \*\*Quality Metrics\*\*:

- `Stats`: 包含各种质量指标，如信号质量、碱基呼叫的置信度等。

7. \*\*Sample Information\*\*:

- `Sample`: 包含样本的描述信息，如样本ID、制备方法等。

8. \*\*Channel Information\*\*:

- `ChannelID`: 包含通道的标识信息，如通道编号、使用的化学品等。

9. \*\*Read Information\*\*:

- `ReadID`: 包含读取的标识信息，如读取编号、读取时间等。

10. \*\*Experiment Information\*\*:

- `Experiment`: 包含实验的详细信息，如实验条件、使用的试剂等。

11. \*\*Configuration\*\*:

- `Config`: 包含测序过程中使用的配置文件信息。

12. \*\*Group\*\*:

- `Group`: 包含组信息，这可能是一个包含多个读取的组。

13. \*\*Tracking\*\*:

- `TrackingID`: 包含追踪信息，用于关联不同的测序运行。

这些子字段的数据可以通过HDF5查看器或者编程接口（如`h5py`）来访问。例如，使用Python的`h5py`库，可以编写脚本来读取和分析这些子字段中的数据。

```python

import h5py

# 打开FAST5文件

with h5py.File('example.fast5', 'r') as f:

# 读取原始数据

raw\_data = f['Raw/Reads/Read\_1'][()]

# 读取基频分析结果

basecall\_1d = f['Basecall\_1D/Read\_1/Model'][()]

# 读取质量指标

stats = f['Stats/Read\_1'][()]

# 读取样本信息

sample\_info = f['Sample'][()]

# 读取通道信息

channel\_info = f['ChannelID'][()]

# 读取读取信息

read\_info = f['ReadID'][()]

# 读取实验信息

experiment\_info = f['Experiment'][()]

# 读取配置信息

config = f['Config'][()]

# 读取组信息

group\_info = f['Group'][()]

# 读取追踪信息

tracking\_info = f['TrackingID'][()]

```

请注意，上述代码只是一个示例，实际的子字段名称可能会根据测序仪的型号和软件版本有所不同。在处理FAST5文件时，通常需要参考特定测序平台和软件的文档来了解具体的数据结构。