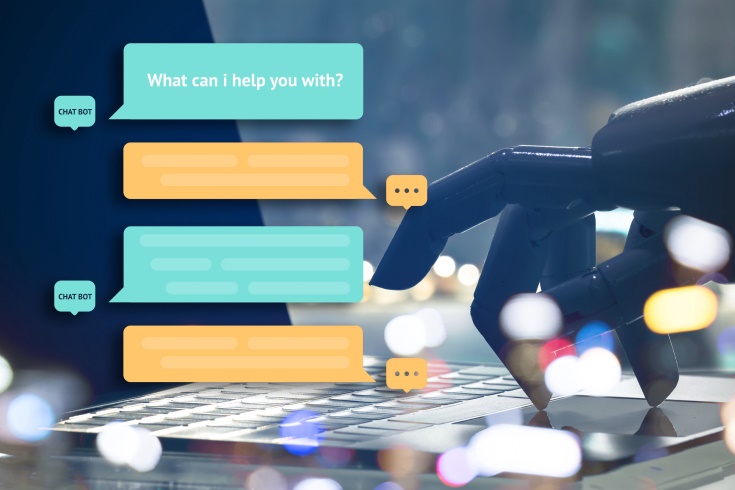
**Spark - Chat intent and most reccuring issue identification use case (use Dataframe or SQL or Both), try with bath and streaming by pasting the chat in the socket of kafka console producer and have the stopwords in the filesystem. Note: Dataset is attached in the last page of this document**



**Data Flow Architecture**

Lookup

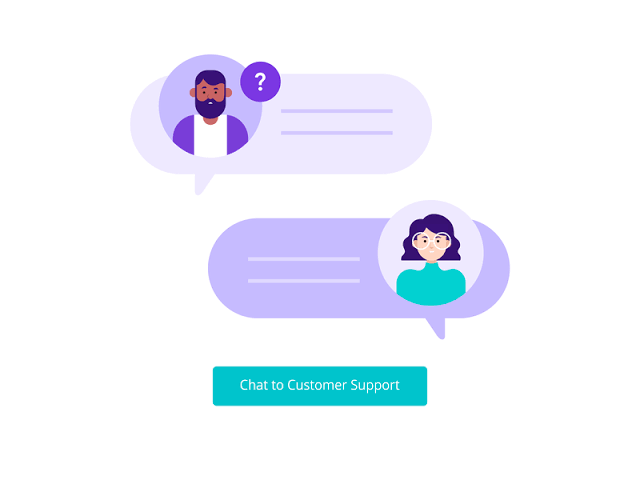
**Stop Words File**

Spark /Spark Streaming

Extract

Join, Enrich, Convert, Pivot etc.

**File System / Socket / kafka**



Aggregation

Push

Storage

**{Json Format}**

Hive

**Try the below scenario with File source then try with socket or kafka streaming of the chat data and stopwords data in the filesystem.**

1. Load the chat data into a Dataframe using ~ delimiter.

2. Define the column names as id,chat,'type'

3. Filter only the records contains 'type' as 'c' (which is the customer interactions)

4. Remove the column 'type' from the above dataframe, hence the resultant dataframe contains only id and chat and convert to tempview.

5. Use SQL split function on the 'chat' column with the delimiter as ' ' (space) for eg. if a chat looks like this (i have internet issue from yesterday) and convert the chat column as array type and name it as chat\_split which has to looks like [i,have,internet,issue,from,yesterday]

6. Use SQL explode function to pivot for example if we have splitted chat data looks like this

id,chat\_split

1,[i,have,internet,issue,from,yesterday]

then the exploded data looks like below

1,i

1,have

1,internet

1,issue

1,from

1,yesterday

7. Download the stopwords data from online used for removing the unwanted additional text we use when chatting.

for eg. A single column contains values such as (a,above,accross,after,an,as,at) etc.

8. Load the stopwords into dataframe with the column name stopword and convert to tempview.

9. Write a left outer join between chat tempview and stopwords tempview and filter all nulls (or) use subquery with not in option to filter all stop words from the actual chat tempview using the stopwords tempview created above.

10. Load the final result into a hive table should have only result as given below using append option.

**eg.**

**Input data in the chat file or socket stream:**

1~hi~c

100~hello morning, how can i help you today~a

1~i have issues in my tv connection which is notworking~c

100~I m really sorry about that. Tell me what happened~a

1~heavy rain in the morning caused my stb and my tv is notworking~c

100~let me take your request, hopefully your issue will be resolved soon~a

**Output data into hive table:**

1,issues

1,tv

1,connection

1,notworking

1,heavy

1,rain

1,morning

1,caused

1,stb

1,tv

1,notworking

11. Identify the most recurring keywords used by the customer in all the chats by grouping based on the keywords used with count of keywords. use group by and count functions in the sql

for eg.

issues,2

notworking,2

tv,2

12. Store the above result in a **json** format with column names as chatkeywords, occurance

**DataSet:**

