**BOOKIPI PySpark Transformation Integration**

1. Prepare pyspark script uploaded (BillingPeriod\_Transformation.py)
2. Prepare python environment to install Google Cloud SDK and Apace Beam in local environment – Needed prerequisites to run google pipeline

* **Apache-Beam**
* **Google-Cloud-BigQuery**
* **Pymongo**
* **Pandas**

1. Prepare python script uploaded (Bookipi\_Python\_IngestData.py)
2. Upload python script into Google Cloud Storage (GCS)   
   - *gsutil cp Bookipi\_Python\_IngestData.py gs://Bookipi\_Bucket/* *Bookipi\_Python\_IngestData.py*
3. Submit Dataflow Job

* *gcloud dataflow jobs run Bookipi\_Pipeline \ --gcs-location gs:// Bookipi\_Bucket /* *Bookipi\_Python\_IngestData.py \ --* *us-central1 \*

1. Prepare error handling python script (Bookipi\_Python\_ErrorHandling.py)
2. Prepare BigQuery Stored Procedure to handle additional data cleanup (Bookipi\_BigQuery\_Cleanup\_Staging.sql)
3. Prepare BigQuery Stored Procedure to handle additional data cleanup (Bookipi\_BigQuery\_Cleanup\_Production.sql)
4. Schedule job via Google Cloud Scheduler – Change Frequency depending on MongoDB data refresh