Hillsboro Python Machine Learning Meetup

Sep/2019

Ernest Bonat, Ph.D.

Senior Software Engineer

Senior Data Scientist

DAT Wi-Fi

Username: DAT Guest

Password: beaverton dat

- 6:00 – 6:40 pm: Pizza, **water only** and networking.

- 6:40 – 6:45 pm: Welcome message by Ernest Bonat, Ph.D.

- 6:45 – 8:00 pm: Presentation and open discussions.

- 8.00 pm – 9.00 pm: Coding and learning session. Bring your Python development laptop!

Why did I create this meetup?

1. Bad traffic to Portland downtown.

2. Very hard to find a parking lot.

3. Bad Python presentation code and old used Python tools.

Our Meetup Mission:

1. *"Come, Listen, Code and Learn"*

2. Finding and presenting best practices of Machine Learning using Python Data Ecosystem.

3. Create great networking place for Hillsboro-Beaverton Data Scientists.

Today Presentation

“Best Practices of Extract-Transform-Load (ETL) Packages Design and Development”

Ernest Bonat, Ph.D.

Senior Data Scientist

From Machine Learning – **Data Preprocessing** (60% - 80% of the whole work!)

"Without clean data your talk may look stupid"

Ernest Bonat, Ph.D.

Machine Learning Work Flow:

1. Data Load
2. Data Exploration
3. Data Visualization
4. **Data Preprocessing**
5. Models Train and Validation
6. Models Test
7. Model Deployment
8. Making Business Decisions…

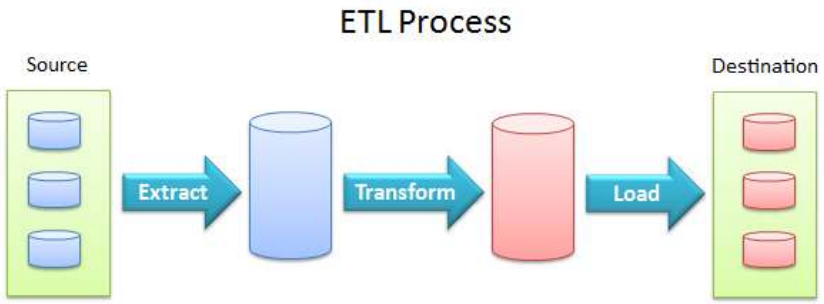
Data Preprocessing – using ETL process:

Let's define the ETL process very simple and clear.

1. **Extract -** data extraction from one or many sources.
2. **Transform -** data transformation (preprocessing or cleansing). **A very clear data transformation requirements document is required.**
3. **Load -** data loading to one or many sources.

Data Sources: database engines (Oracle, SQL Server, PostgreSQL, MySQL, SQLite, MongoDB, MariaDB, etc.); data files (LOG, CSV, XML, XLS, JSON, etc.), any possible data sources and/or their combinations, etc.

Define ETL Process (package)



Similar 3-tier App Architecture = 3-layer Extract - Transform - Load)

3-Tier Architecture	ETL Process
Presentation layer ⬇ ⬆	Load Layer ⬆
Business Logic layer ⬇ ⬆	Transform layer ⬆
Data Access Layer	Extract Layer

Project Folder Structure

Folder Name	Folder Definition
config	application configuration file (app.cfg)
data	data source files
extract	extract files
library	general library files
load	load files
log	application log file (app.log)
transform	transform files

Compile to EXE file using PyInstaller (VERY IMPORTANT!)

pyinstaller --onedir --name=installer_name windowed "C:\python_file_path\main_file_name.py"

EXE File and Folder Deployment (“dist” folder will be created)

1. Task Schedule
2. Windows Services
3. ?

Final blog paper by the end of October/2019

“ELT Process Design and Development with 3-Tier Architecture Pattern”