Process Mining With Jupyter Notebook And Disco

Faruque Ahmed
fahmed.max@gmail.com

Python have the best quality of documentation.

The maturity and stability of the fundamental numerical libraries are well known.

pandas is a Python library of data structures and statistical tools.

Jupyter is the GUI tool set that widely use for the Python open source community for Data pre-processing.

Faruque Ahmed
fahmed.max@gmail.com

The Data Set - **BPI Challenge 2019**

The data was recorded for the execution of **Purchase-To-Pay processes**.

- It was recorded possibly from an ERP system.
- The log contains 1,595,923 events, with 22 columns.
- Character set - Windows-1252 or CP-1252

1595923 rows × 22 columns

```
df = pd.read_csv("BPI2019.csv", encoding='cp1252')
```

Faruque Ahmed
fahmed.max@gmail.com

Discover : Case ID

```
 #   Column                          Non-Null Count     Dtype
---  ------                          --------------     -----
 0   eventID                         1595923 non-null   int64
```

```
In [6]: df['eventID '].unique()

Out[6]: array([  65781719105536,   65777424138241,   65777424138240, ...,
         1009441868611588,  273679611068428,   77635828842576], dtype=int64)

In [7]: pd.value_counts(df['eventID '])

Out[7]: 697502688870404     1
        1058967136501761    1
        610233248382980     1
        211767657496581     1
        77249281785859      1
                           ..
        214318868070410     1
        796652008898560     1
        911705592823808     1
        682143885819904     1
        0                   1
        Name: eventID , Length: 1595923, dty

In [8]: count = df['eventID '].value_counts()

In [9]: (count == 1).value_counts()

Out[9]: True    1595923
        Name: eventID , dtype: int64
```

```
In [10]: df.loc[df['eventID ']== 697502688870404,:]

Out[10]:
```

	eventID	case Spend area text	case Company	case Document Type	case Sub spend area text	case Purchasing Document	case Purch. Doc. Category name	case Vendor	case Item Type	case Item Category	...	case Name	case GR-Based Inv. Verif.	ca It
1505633	697502688870404	Packaging	companyID_0000	Standard PO	Labels	4508048579	Purchase order	vendorID_0120	Standard	3-way match, invoice before GR	...	vendor_0119	False	

1 rows × 22 columns

eventID can not be a **case ID** for Process Discovery

Faruque Ahmed
fahmed.max@gmail.com

Discover : Case ID

```
 #   Column                          Non-Null Count    Dtype
---  ------                          --------------    -----
 0   eventID                         1595923 non-null  int64
 1   case Spend area text            1579629 non-null  object
 2   case Company                    1595923 non-null  object
 3   case Document Type              1595923 non-null  object
 4   case Sub spend area text        1579629 non-null  object
 5   case Purchasing Document        1595923 non-null  int64
 6   case Purch. Doc. Category name  1595923 non-null  object
 7   case Vendor                     1595923 non-null  object
 8   case Item Type                  1595923 non-null  object
```

case Purchasing Document

- No null value
- Int64
- But no flow **Activity**

case concept:name Is the perfect match

- Same PO
- Good flow of Activity
- Same Amount (EUR)
- Continuous timestamp

case concept:name	case Goods Receipt	event User	event org:resource	event concept:name	event Cumulative net worth (EUR)	event time:timestamp
4508048579_00130	True	user_057	user_057	Create Purchase Order Item	418.0	10-08-2018 16:20:00.000
4508048579_00130	True	user_034	user_034	Record Goods Receipt	418.0	29-08-2018 11:17:00.000
4508048579_00130	True	NONE	NONE	Vendor creates invoice	418.0	30-08-2018 23:59:00.000
4508048579_00130	True	user_012	user_012	Record Invoice Receipt	418.0	31-08-2018 12:26:00.000
4508048579_00130	True	user_023	user_023	Remove Payment Block	418.0	13-12-2018 08:03:00.000
4508048579_00130	True	user_002	user_002	Clear Invoice	418.0	13-12-2018 14:08:00.000

```
In [19]: case_id = df['case concept:name'].replace('_', '', regex=True)
```

```
In [23]: df['case concept:name'] = pd.to_numeric(df['case concept:name'])
```

Faruque Ahmed
fahmed.max@gmail.com

event time:timestamp

```
In [14]: pd.to_datetime(df['event time:timestamp']).head()

Out[14]: 0   1948-01-26 23:59:00
         1   1948-01-26 23:59:00
         2   1948-01-26 23:59:00
         3   1948-01-26 23:59:00
         4   1948-01-26 23:59:00
         Name: event time:timestamp, dtype: datetime64[ns]
```

It was a Object type data.

For the search and manipulation we change it to **datetime64**

- Purchase orders submitted mostly in 2018.
- Some purchase orders start from 1948 (only 295 out of 1,595,923)
- These 295 purchase orders are outlier
- We delete those process

```
In [27]: df.loc[df['timestamp'] < '2017-12-30 00:00:00','timestamp']

Out[27]: 0     1948-01-26 23:59:00
         1     1948-01-26 23:59:00
         2     1948-01-26 23:59:00
         3     1948-01-26 23:59:00
         4     1948-01-26 23:59:00
                      ...
         290   2017-12-29 23:59:00
         291   2017-12-29 23:59:00
         292   2017-12-29 23:59:00
         293   2017-12-29 23:59:00
         294   2017-12-29 23:59:00
         Name: timestamp, Length: 295, dtype: datetime64[ns]
```

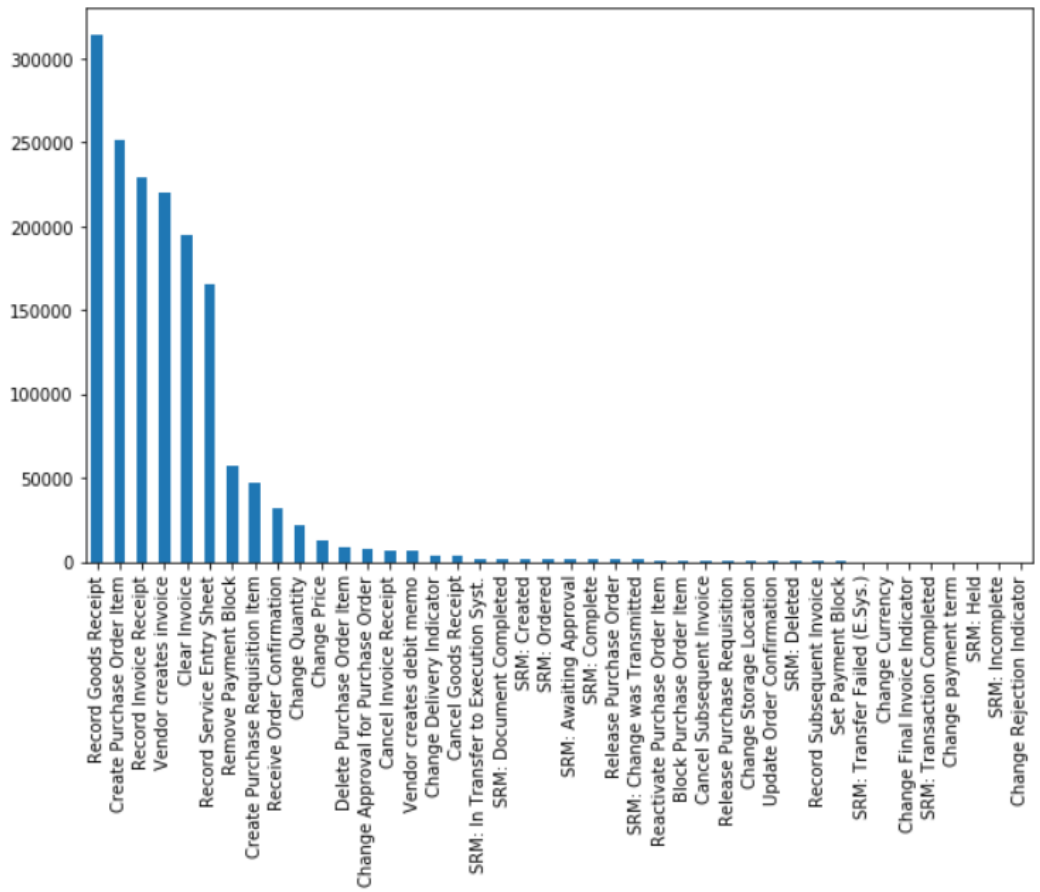
```
In [25]: df['timestamp'].dt.year.value_counts()

Out[25]: 2018    1550468
         2019      45135
         2017        223
         2008         45
         2001         22
         1948         10
         1993          9
         2016          6
         2015          3
         2020          2
         Name: timestamp, dtype: int64
```

Faruque Ahmed
fahmed.max@gmail.com

Discover : Activity event concept:name

Activity	Frequency
Record Goods Receipt	314097
Create Purchase Order Item	251734
Record Invoice Receipt	228760
Vendor creates invoice	219919
Clear Invoice	194393
Record Service Entry Sheet	164975
Remove Payment Block	57136
Create Purchase Requisition Item	46592
Receive Order Confirmation	32065
Change Quantity	21449
Change Price	12423
Delete Purchase Order Item	8875
Change Approval for Purchase Order	7541
Cancel Invoice Receipt	7096
Vendor creates debit memo	6255
Change Delivery Indicator	3289
Cancel Goods Receipt	3096

```
In [7]: df['event concept:name'].value_counts().plot.bar(figsize=(10,6))
Out[7]: <matplotlib.axes._subplots.AxesSubplot at 0x232a0c56588>
```

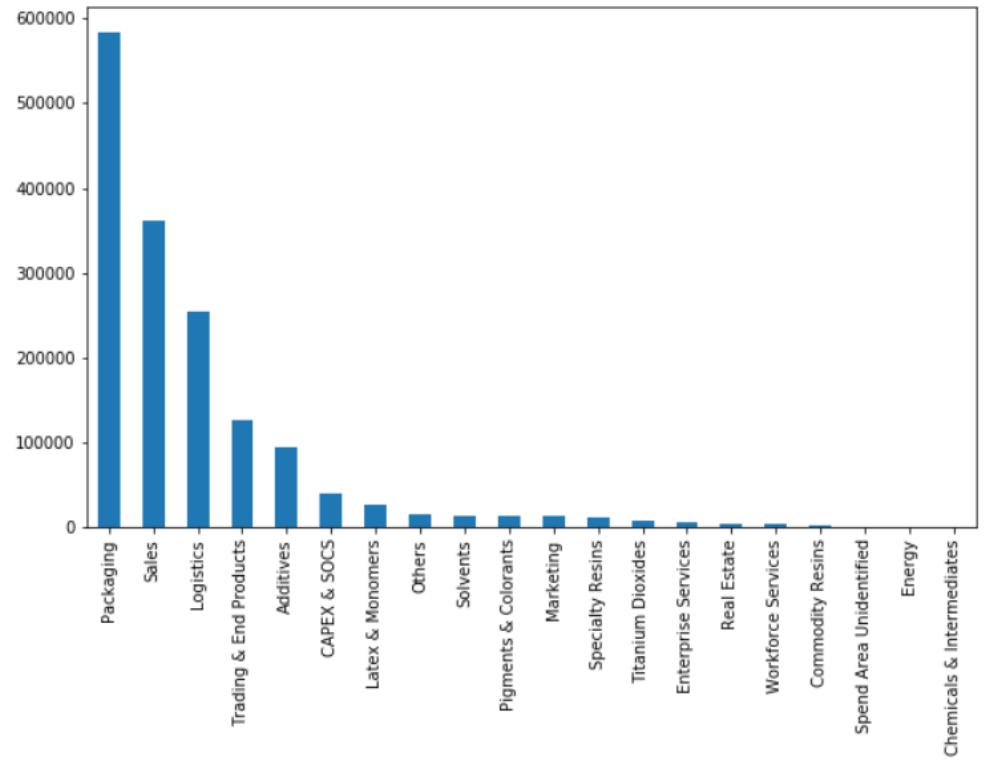


Faruque Ahmed
fahmed.max@gmail.com

Discover : Activity case Spend area text

Activity	Frequency
Packaging	583981
Sales	360774
Logistics	253565
Trading & End Products	126756
Additives	95499
CAPEX & SOCS	40074
Latex & Monomers	27007
Others	15419
Solvents	13889
Pigments & Colorants	13811
Marketing	12994
Specialty Resins	12469
Titanium Dioxides	7453
Enterprise Services	5957
Real Estate	3824
Workforce Services	3441
Commodity Resins	2374

```
In [14]: df['case Spend area text'].value_counts().plot.bar(figsize=(10,6))
Out[14]: <matplotlib.axes._subplots.AxesSubplot at 0x232a1cc4808>
```



Faruque Ahmed
fahmed.max@gmail.com

Discover : User

```
In [7]: print('====case Vendor===')
        print(df['case Vendor'].unique())
        print('')

        print('===case Name====')
        print(df['case Name'].unique())
        print('')

        print('===event User ====')
        print(df['event User'].unique())
        print('')

        print('===event org:resource ====')
        print(df['event org:resource'].unique())
        print('')
```

Four Columns

- case Vendor
- case Name
- event User
- event org:resource

```
====case Vendor===
['vendorID_0670' 'vendorID_0427' 'vendorID_0307' ... 'vendorID_1968'
 'vendorID_1973' 'vendorID_1974']

===case Name====
['vendor_0645' 'vendor_0415' 'vendor_0298' ... 'vendor_1892' 'vendor_1897'
 'vendor_1898']

===event User ====
['NONE' 'user_329' 'user_236' 'user_124' 'batch_03' 'batch_08' 'batch_04'
 'user_033' 'user_036' 'user_038' 'batch_00' 'user_043' 'user_045'
 'user_051' 'user_029' 'user_052' 'user_054' 'user_057' 'user_059'
 'user_060' 'user_064' 'user_066' 'user_068' 'user_070' 'user_072'
 'user_075' 'user_079' 'user_084' 'user_085' 'user_081' 'user_089'
 'user_091' 'user_092' 'user_095' 'user_097' 'user_100' 'user_103'
 'user_105' 'user_108' 'user_000' 'user_110' 'user_116' 'batch_06'
 'batch_02' 'user_154' 'user_086' 'user_113' 'user_118' 'user_122'
```

Faruque Ahmed
fahmed.max@gmail.com

Data quality issues

Data Quality Issues	BPI Challenge 2019
Incorrect Timestamps	X
Missing values	
Missing Events	X
Duplicate Tasks	X
Overlapping Activity Executions	X
Case Heterogeneity	X
Voluminous Data	X
Noisy Data-Outliers	X

Final Data set

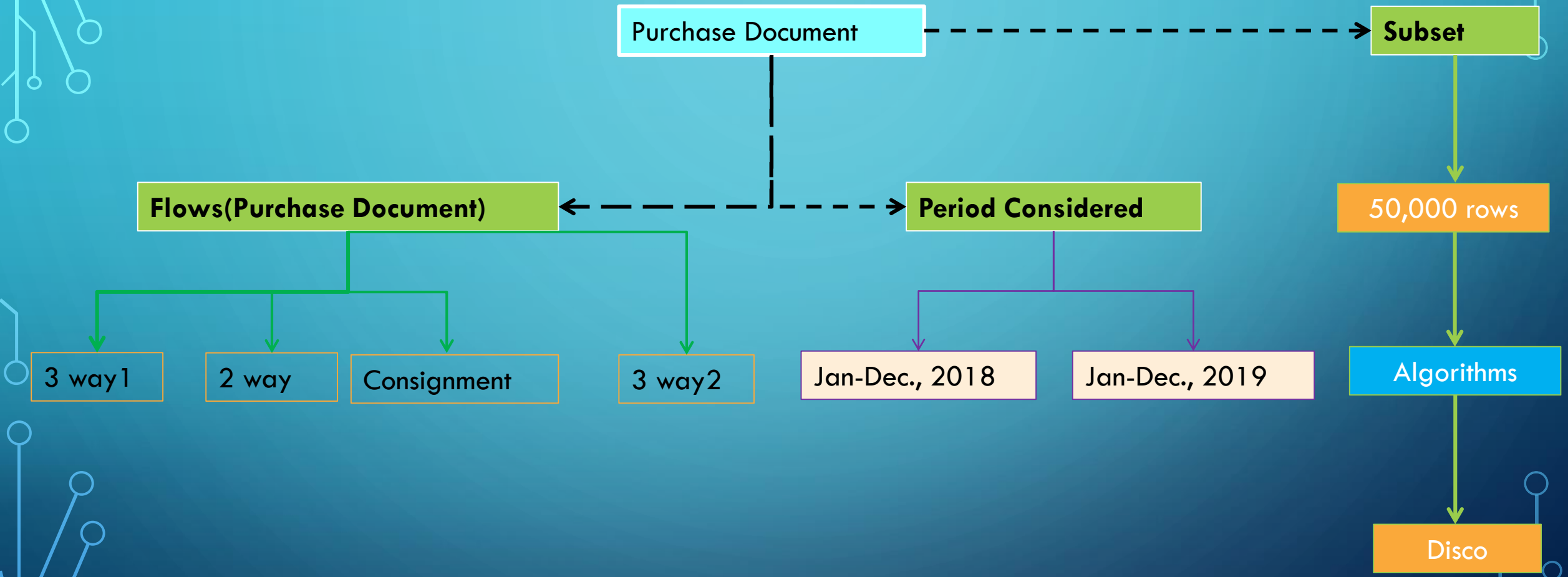
```
In [32]: df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1588570 entries, 0 to 1588569
Data columns (total 13 columns):
 #   Column                          Non-Null Count    Dtype
---  ------                          --------------    -----
 0   case Spend area text            1572311 non-null  object
 1   case Sub spend area text        1572311 non-null  object
 2   case Vendor                     1588570 non-null  object
 3   case Item Type                  1588570 non-null  object
 4   case Item Category              1588570 non-null  object
 5   case Spend classification text  1572311 non-null  object
 6   case Name                       1588570 non-null  object
 7   case GR-Based Inv. Verif.       1588570 non-null  bool
 8   case concept:name               1588570 non-null  int64
 9   event User                      1588570 non-null  object
 10  event concept:name              1588570 non-null  object
 11  event Cumulative net worth (EUR)  1588570 non-null  float64
 12  timestamp                       1588570 non-null  object
dtypes: bool(1), float64(1), int64(1), object(10)
memory usage: 147.0+ MB
```

```
In [27]: df.to_csv('B03_done.csv', encoding='utf-8', index=False)
```

Faruque Ahmed
fahmed.max@gmail.com

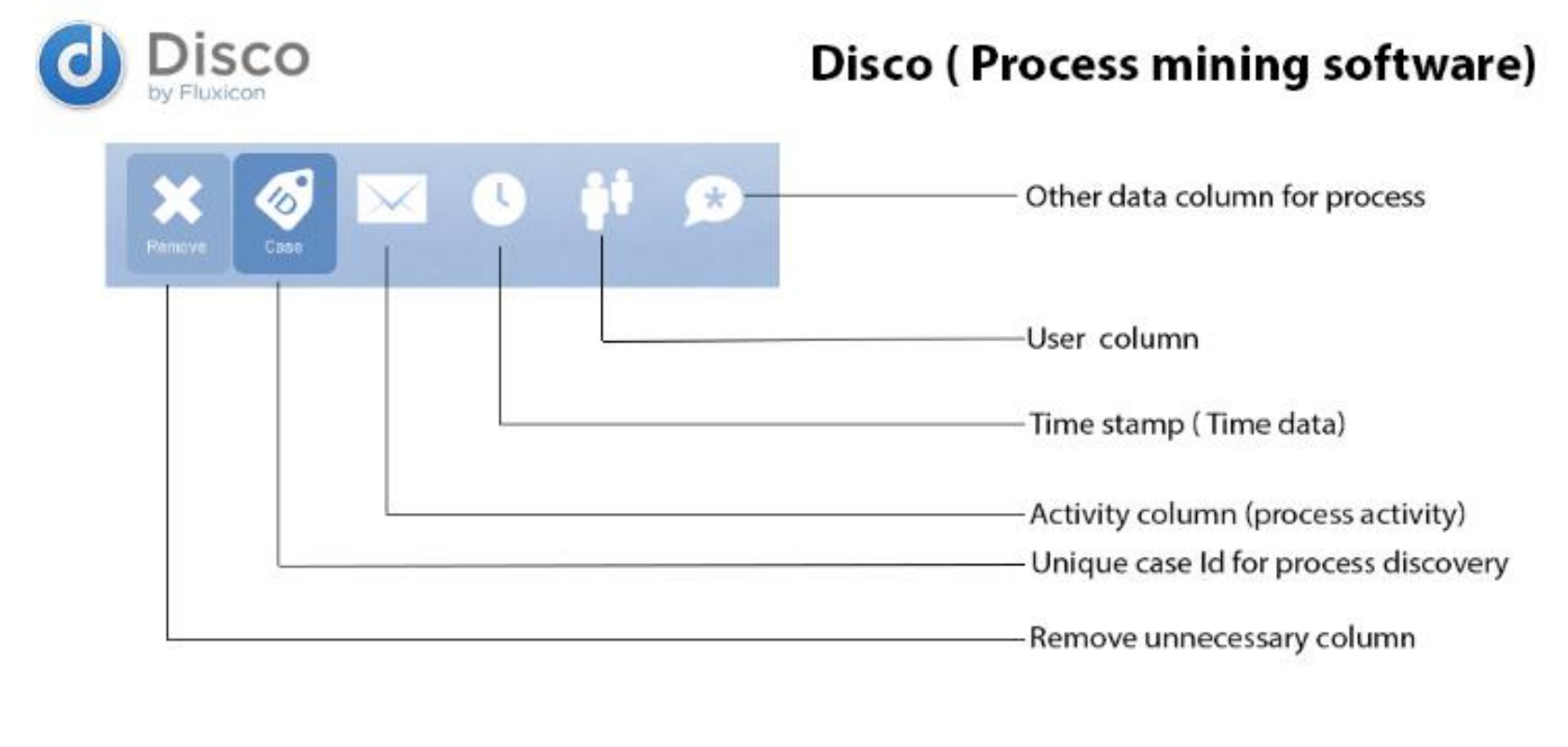
Dataset

Focusing:



Faruque Ahmed
fahmed.max@gmail.com

DISCO Implementation

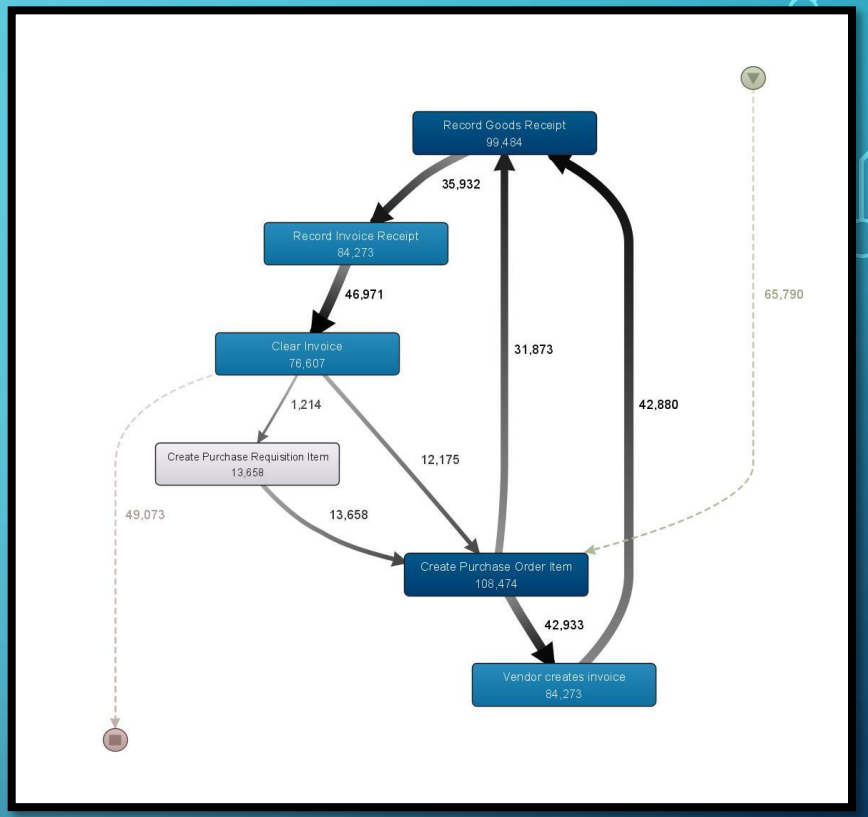


Faruque Ahmed
fahmed.max@gmail.com

The Data Set - **BPI Challenge 2019**

	eventID	case Spend area text	case Company	case Document Type	case Sub spend area text	case Purchasing Document	case Purch. Doc. Category name	case Vendor	case Item Type	case Item Category	...	case Name	case GR-Based Inv. Verif.	case Item	
0	65781719105536	Sales	companyID_0000	Standard PO	Products for Resale	4507004931	Purchase order	vendorID_0670	Standard	3-way match, invoice before GR	...	vendor_0645	False	20	4507
1	65777424138241	Sales	companyID_0000	Standard PO	Products for Resale	4507004931	Purchase order	vendorID_0670	Standard	3-way match, invoice before GR	...	vendor_0645	False	10	4507
2	65777424138240	Sales	companyID_0000	Standard PO	Products for Resale	4507004931	Purchase order	vendorID_0670	Standard	3-way match, invoice before GR	...	vendor_0645	False	10	4507
3	65794604007424	Sales	companyID_0000	Standard PO	Products for Resale	4507004931	Purchase order	vendorID_0670	Standard	3-way match, invoice before GR	...	vendor_0645	False	50	4507
4	65794604007425	Sales	companyID_0000	Standard PO	Products for Resale	4507004931	Purchase order	vendorID_0670	Standard	3-way match, invoice before GR	...	vendor_0645	False	50	4507

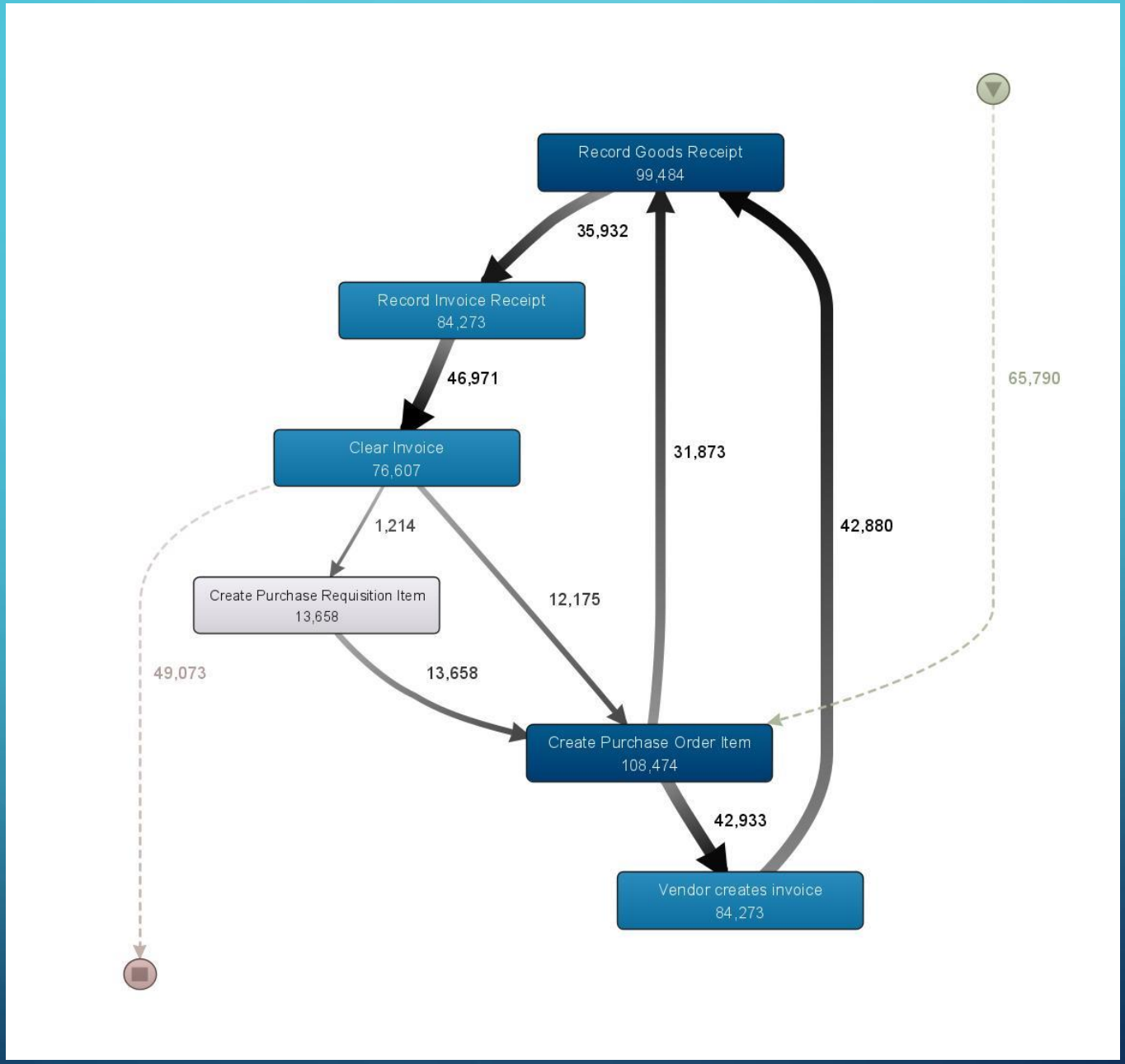
Pre-processing



- Case ID
- Timestamp
- Activity
- Other fields…..

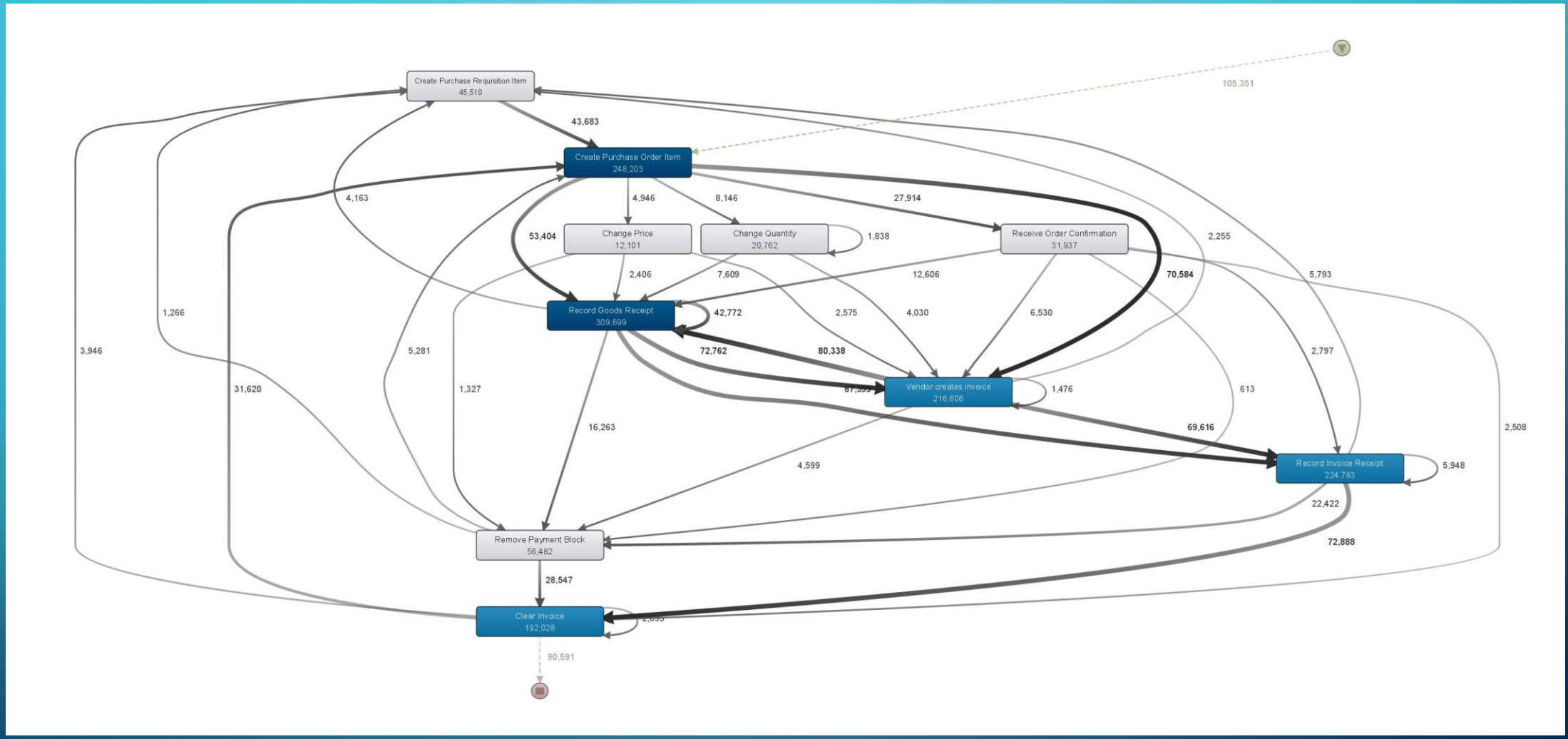
Faruque Ahmed
fahmed.max@gmail.com

Activity



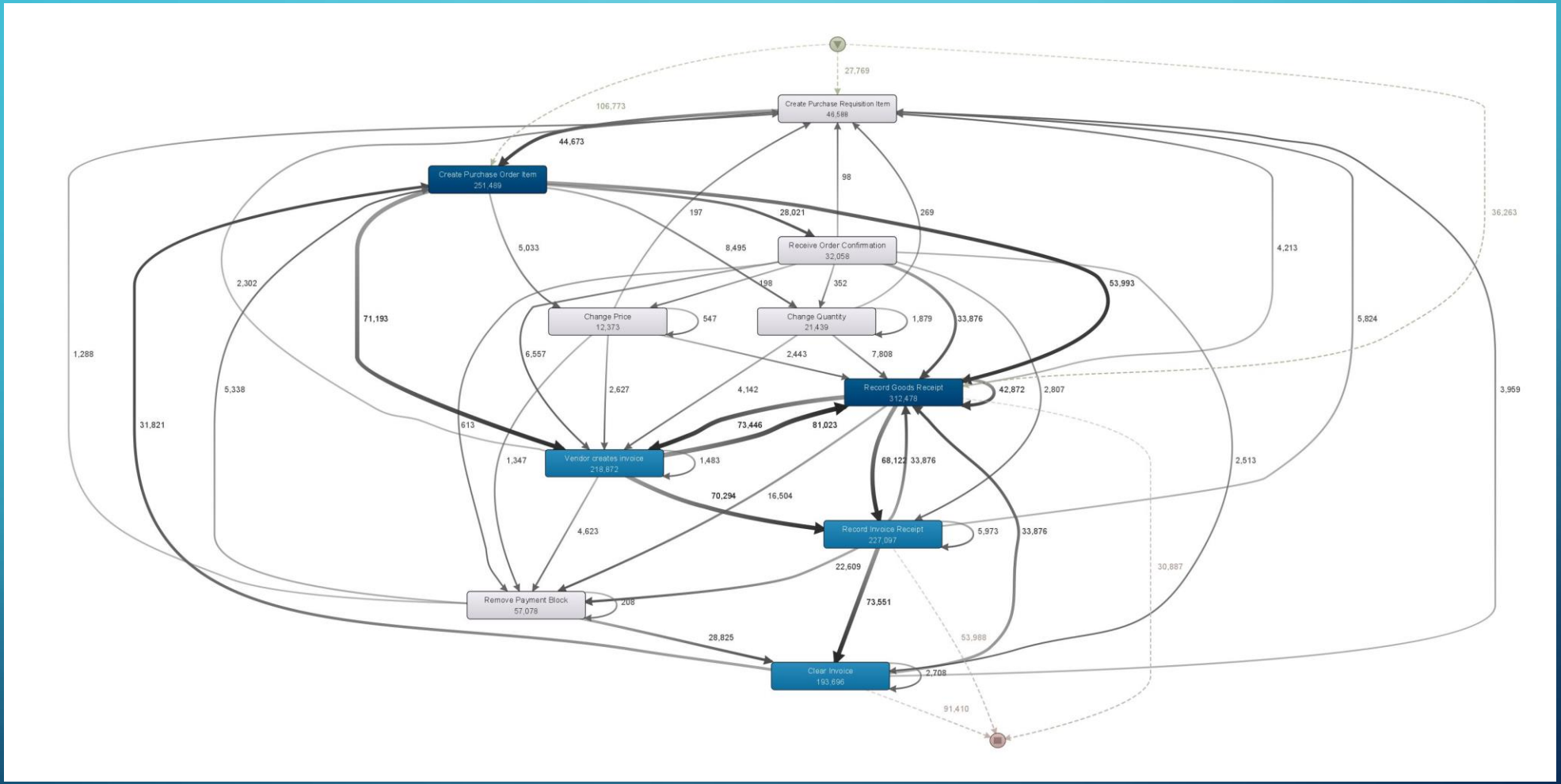
Faruque Ahmed
fahmed.max@gmail.com

Case Spend area text



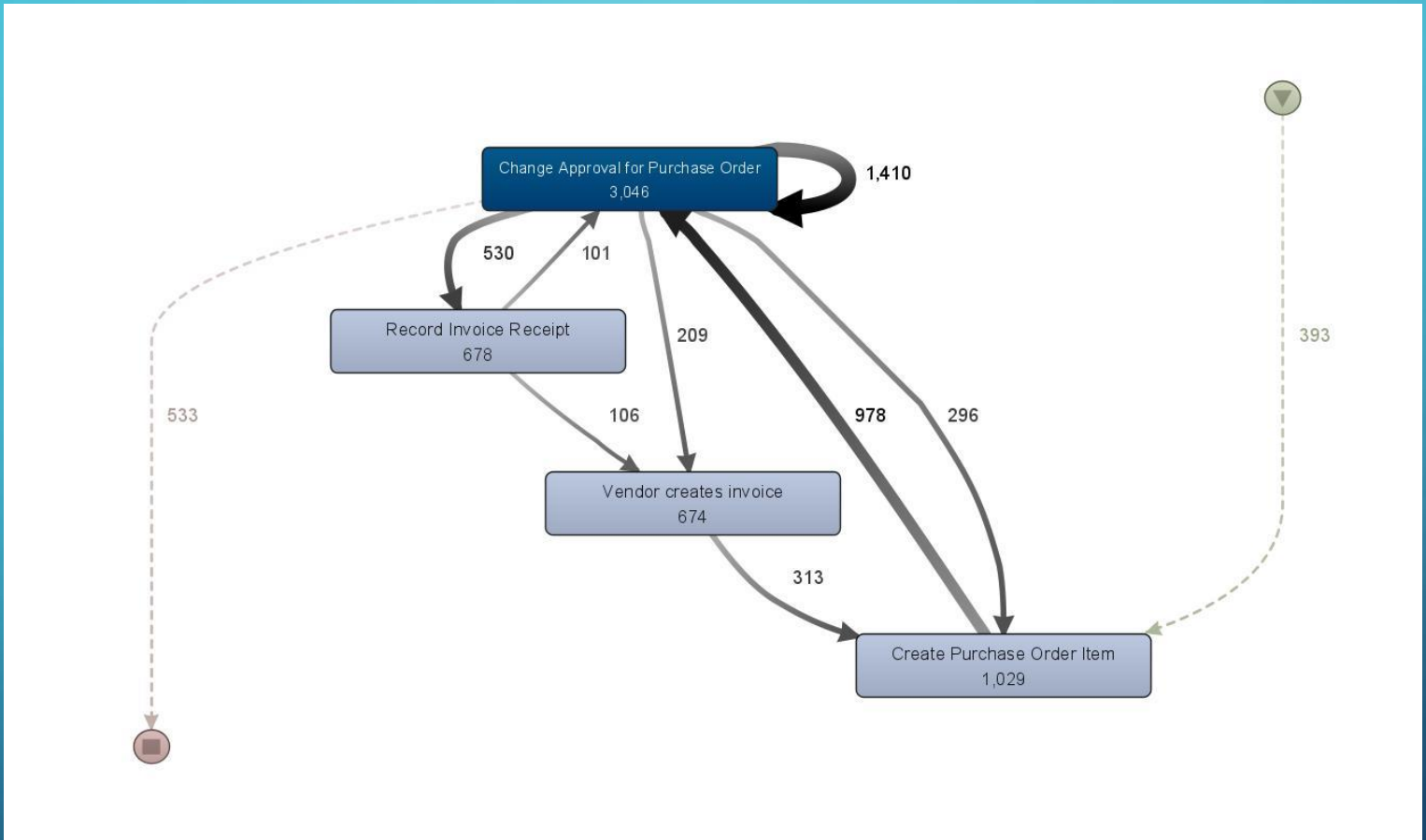
Faruque Ahmed
fahmed.max@gmail.com

Case Sub spend area text



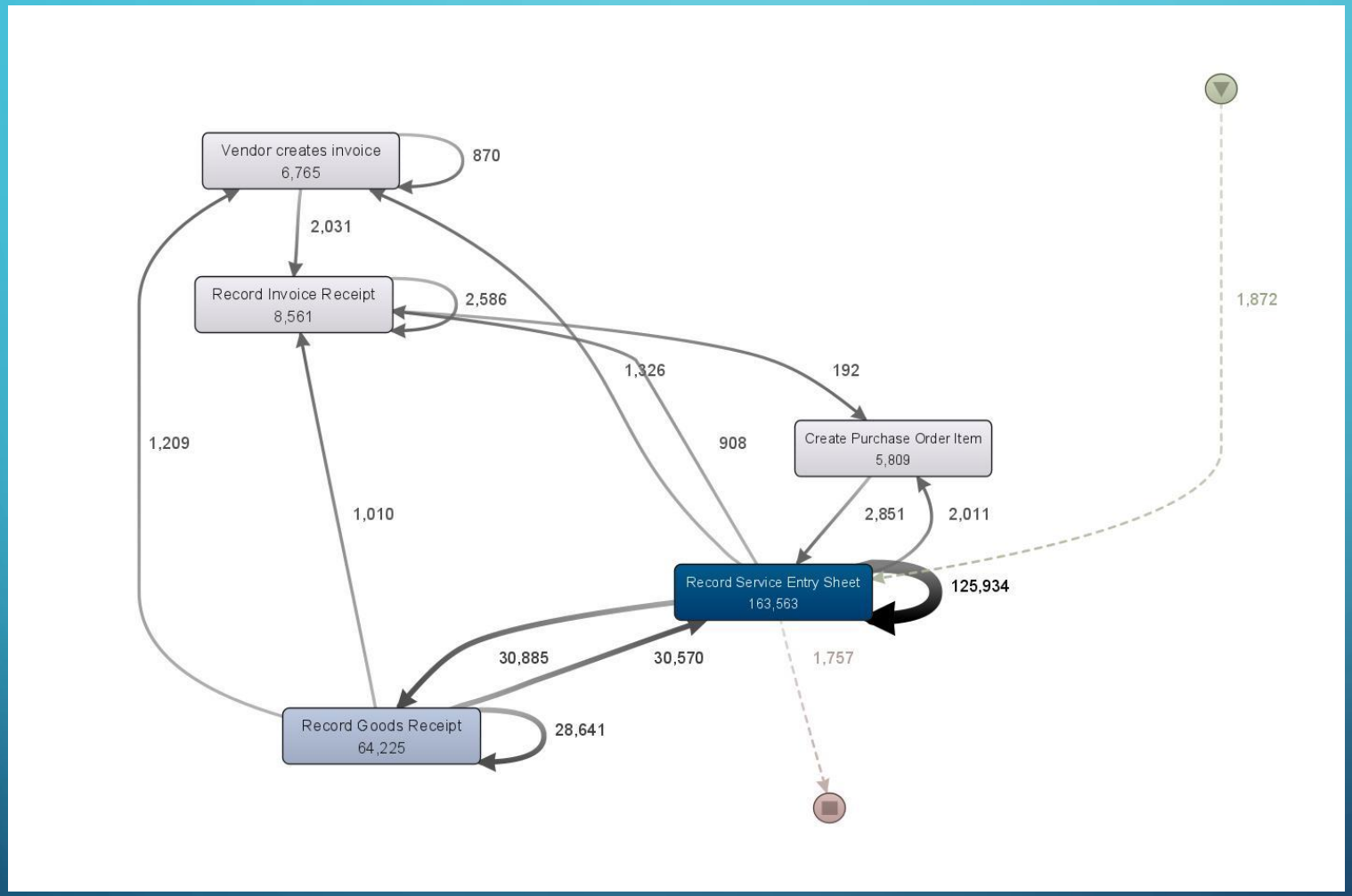
Faruque Ahmed
fahmed.max@gmail.com

Case Item Type - limit



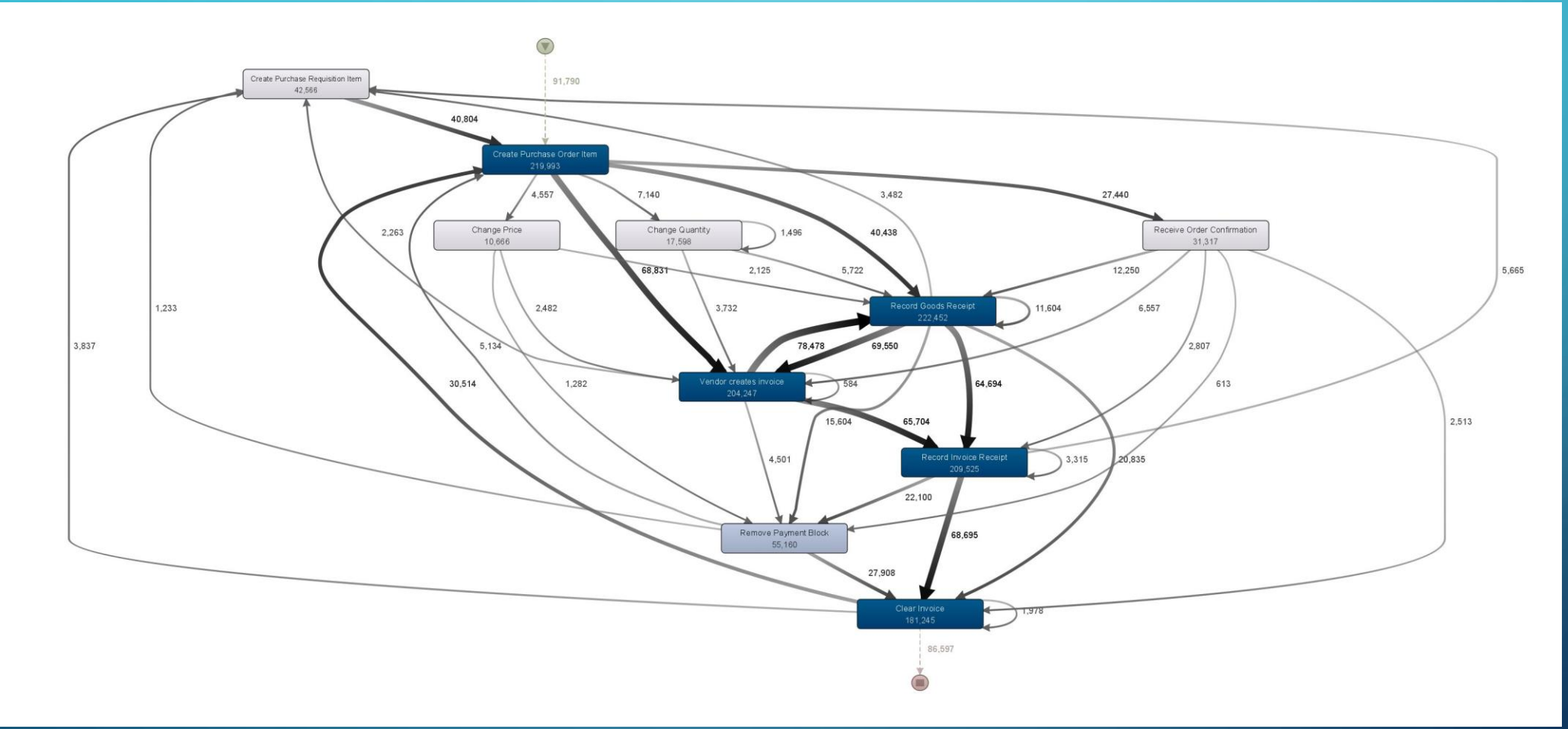
Faruque Ahmed
fahmed.max@gmail.com

Case Item Type - service



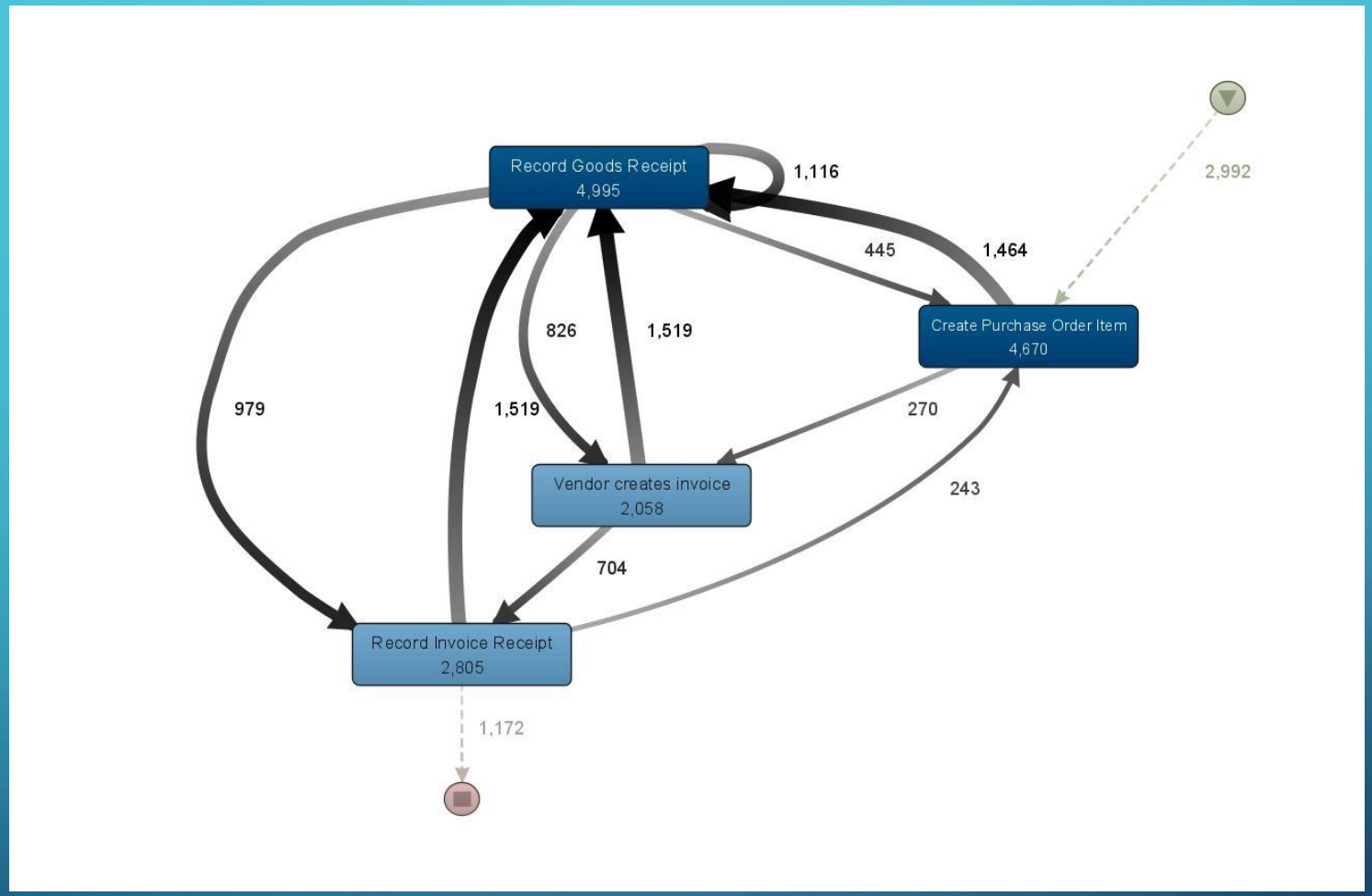
Faruque Ahmed
fahmed.max@gmail.com

Case Item Type - standard



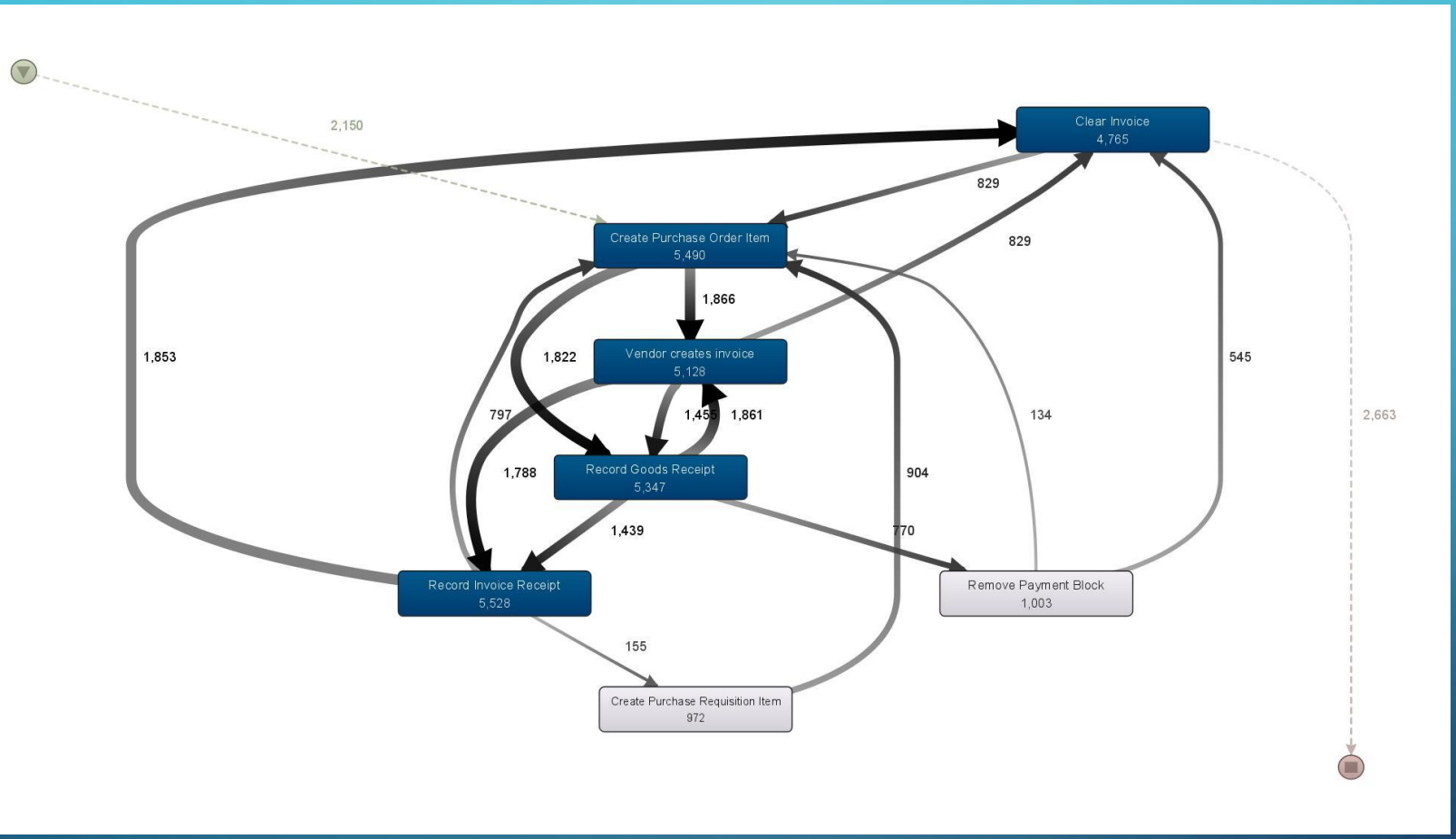
Faruque Ahmed
fahmed.max@gmail.com

Case Item Type - subcontract



Faruque Ahmed
fahmed.max@gmail.com

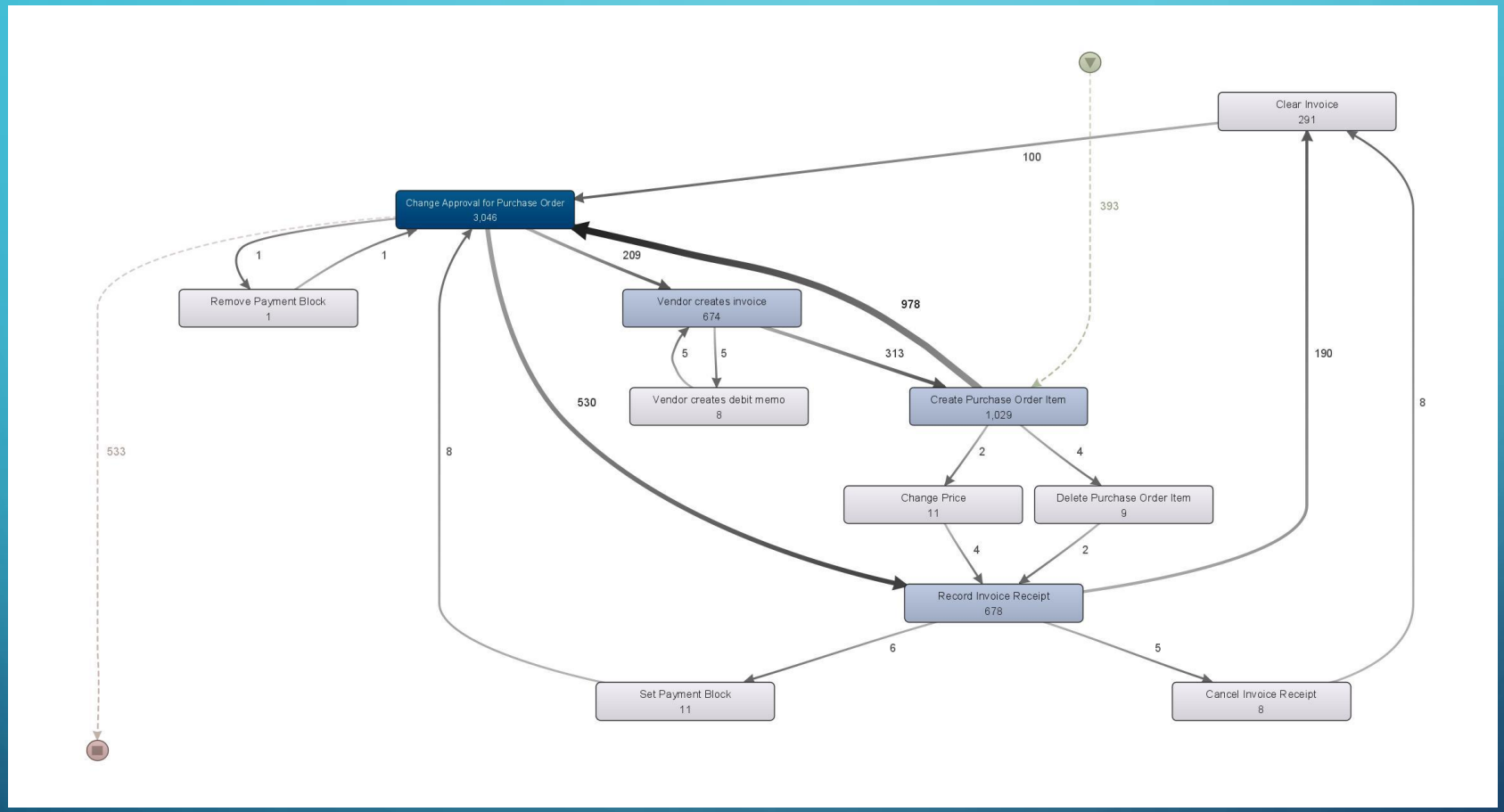
Case Item Type - Third party



Faruque Ahmed
fahmed.max@gmail.com

Case item category – 2-way match

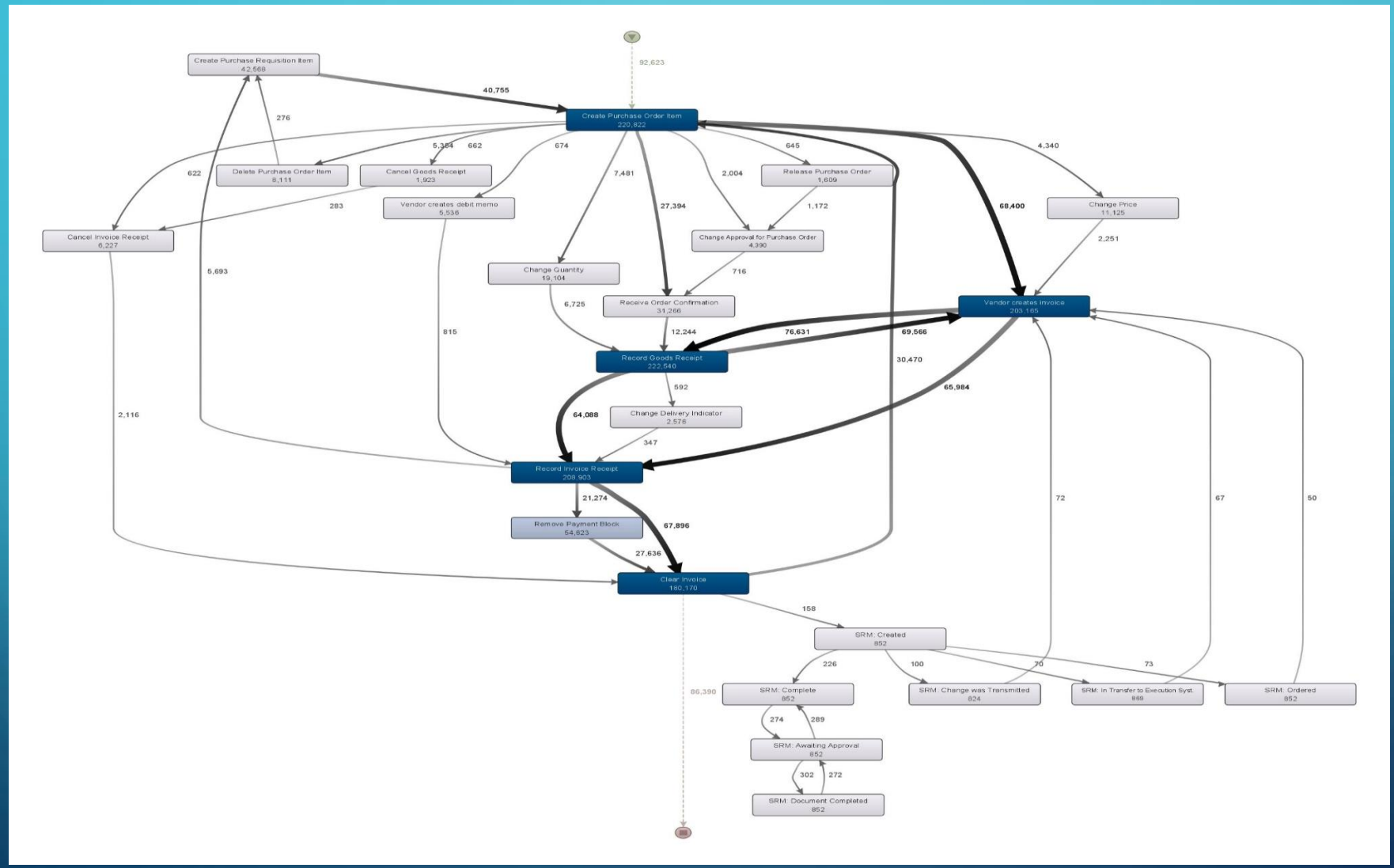
(Values) Invoice ----matches creation (PO)
(Implication) No separate goods receipt message required.
GR-based flag and the Goods Receipt flags set to false



Faruque Ahmed
fahmed.max@gmail.com

Case item category – 3-way match invoice before

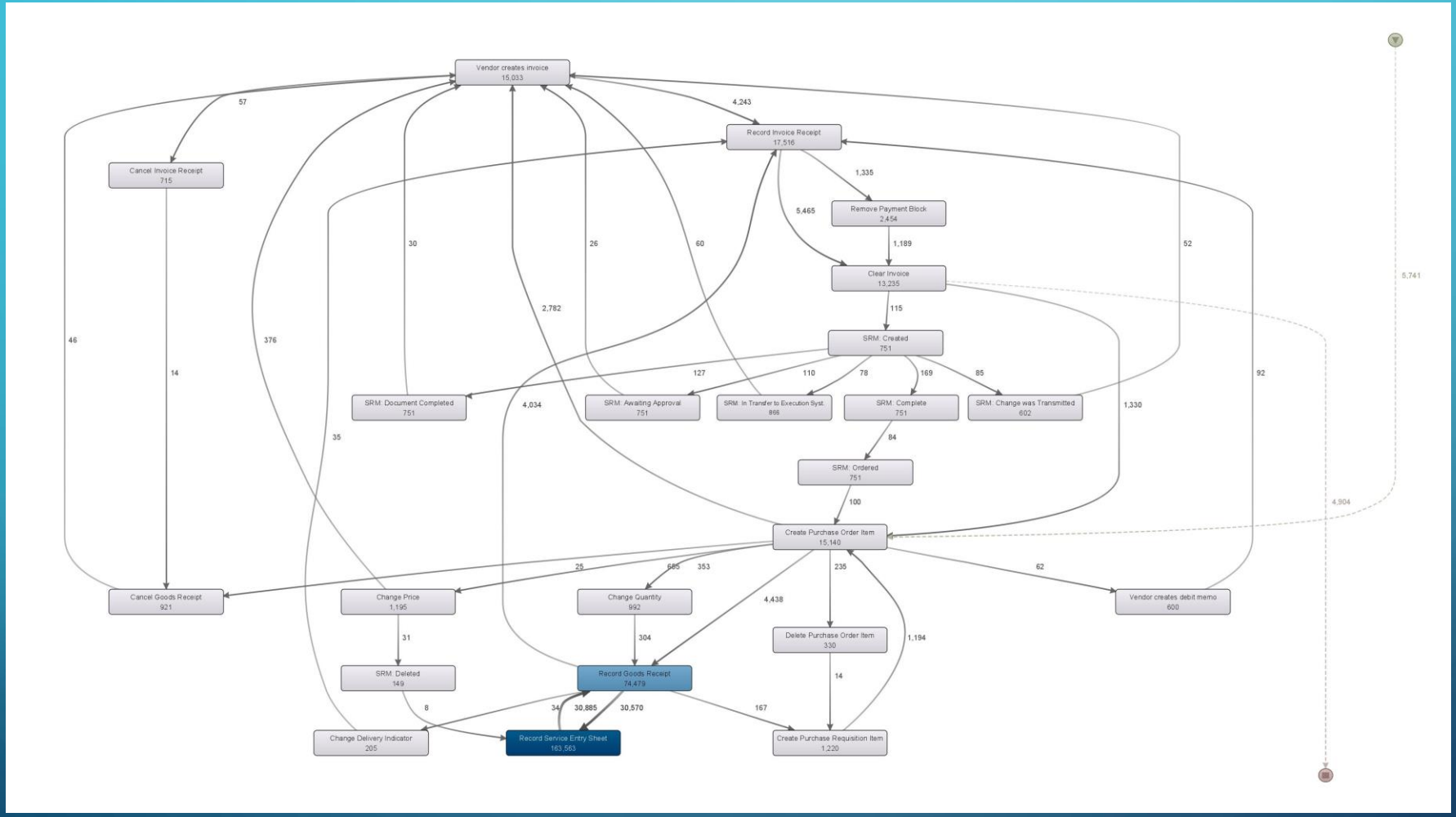
(Values) Purchase Items not requires goods receipt message/GR-based invoicing
(Implication) GR-based IV flag set to false and the Goods Receipt flags set to true.
Invoices entered before the goods are receipt(blocked until goods received
Unblocking is done by a user/a batch process at regular intervals



Faruque Ahmed
fahmed.max@gmail.com

Case item category – 3-way match, invoice after

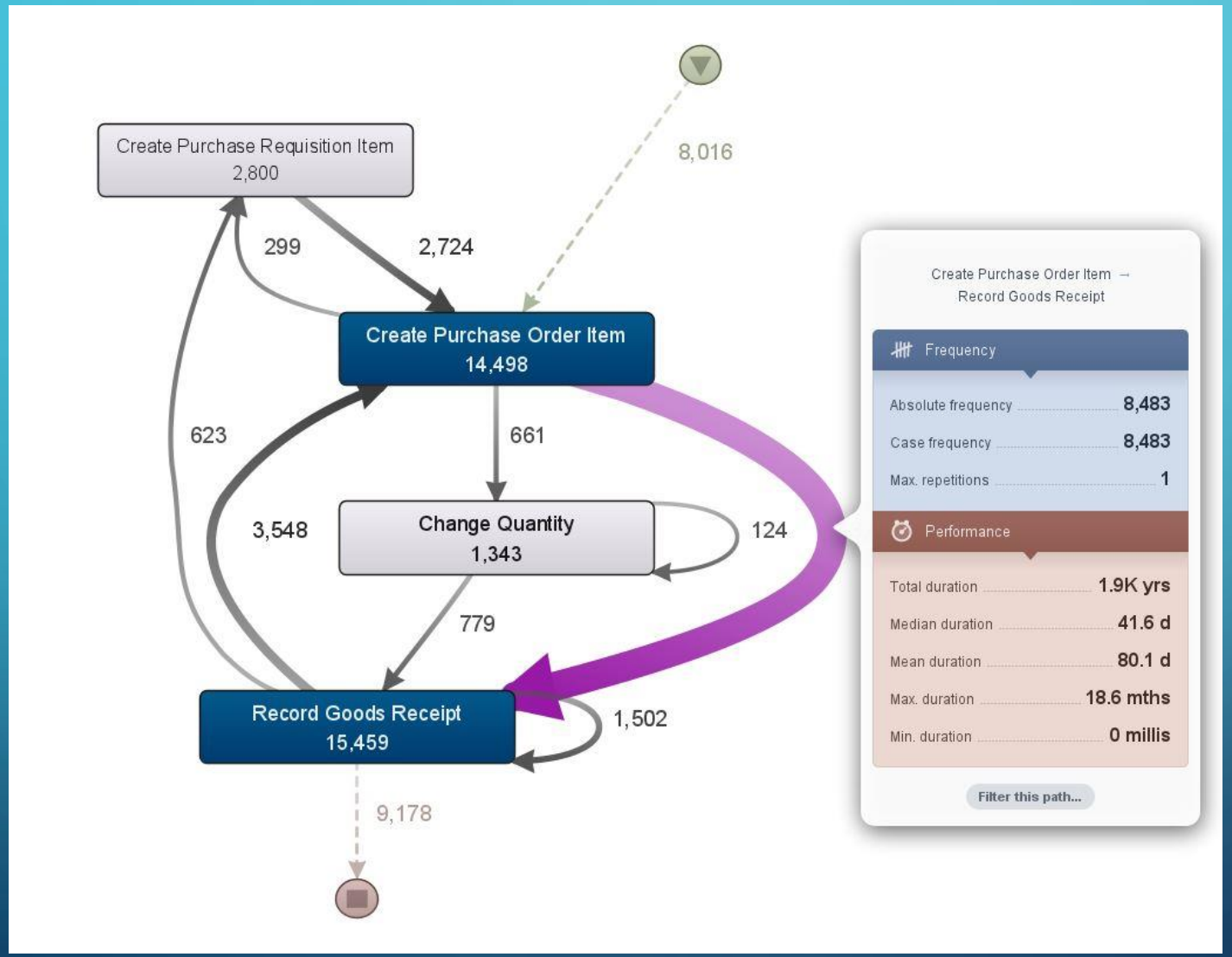
(Values) An invoice receipt message ----matches ------- goods receipt message
(Implication) GR-based flag and the Goods Receipt flags set to true



Faruque Ahmed
fahmed.max@gmail.com

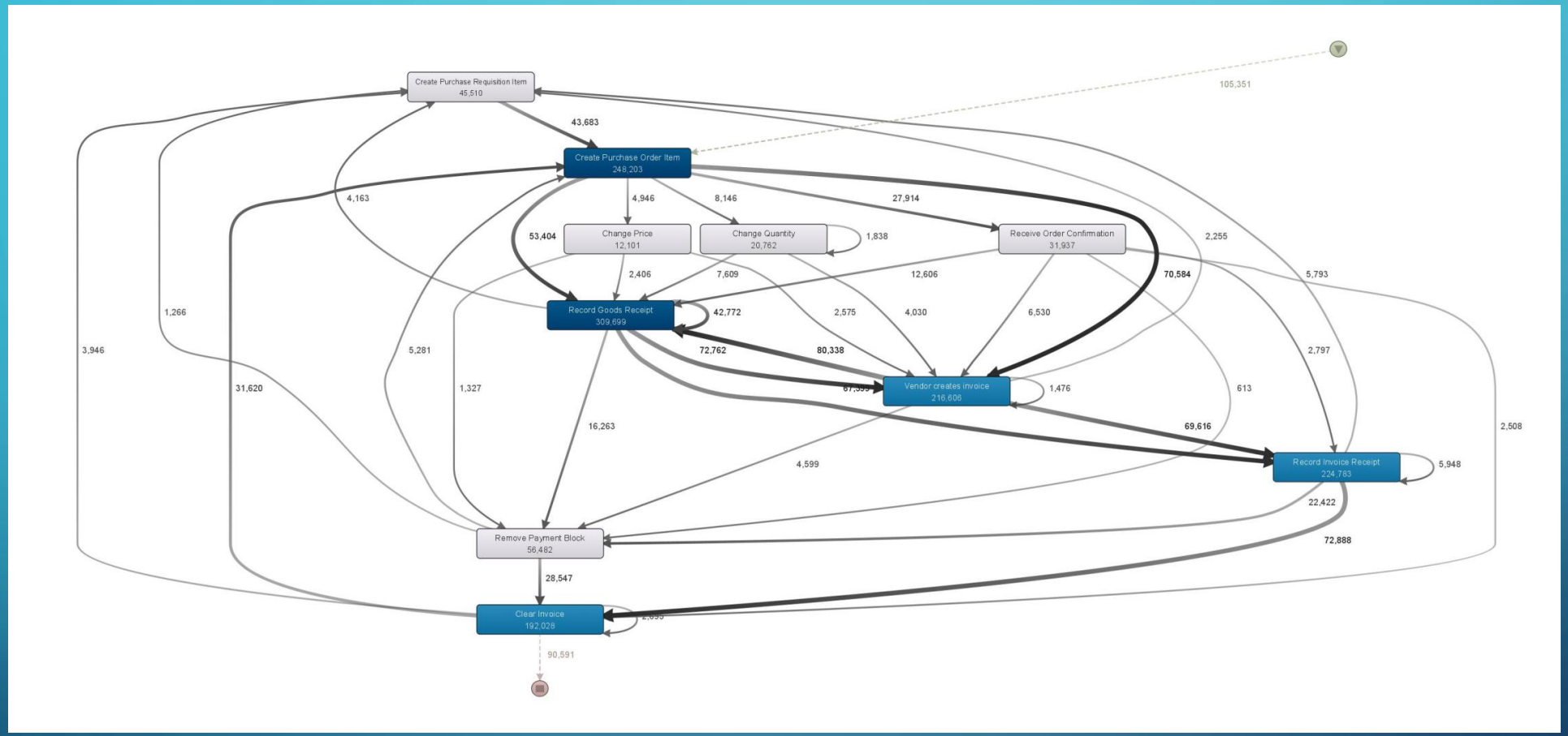
Case Item Type - Consignment

(Values) classification text is PR
(Implication) GR indicator is set to False



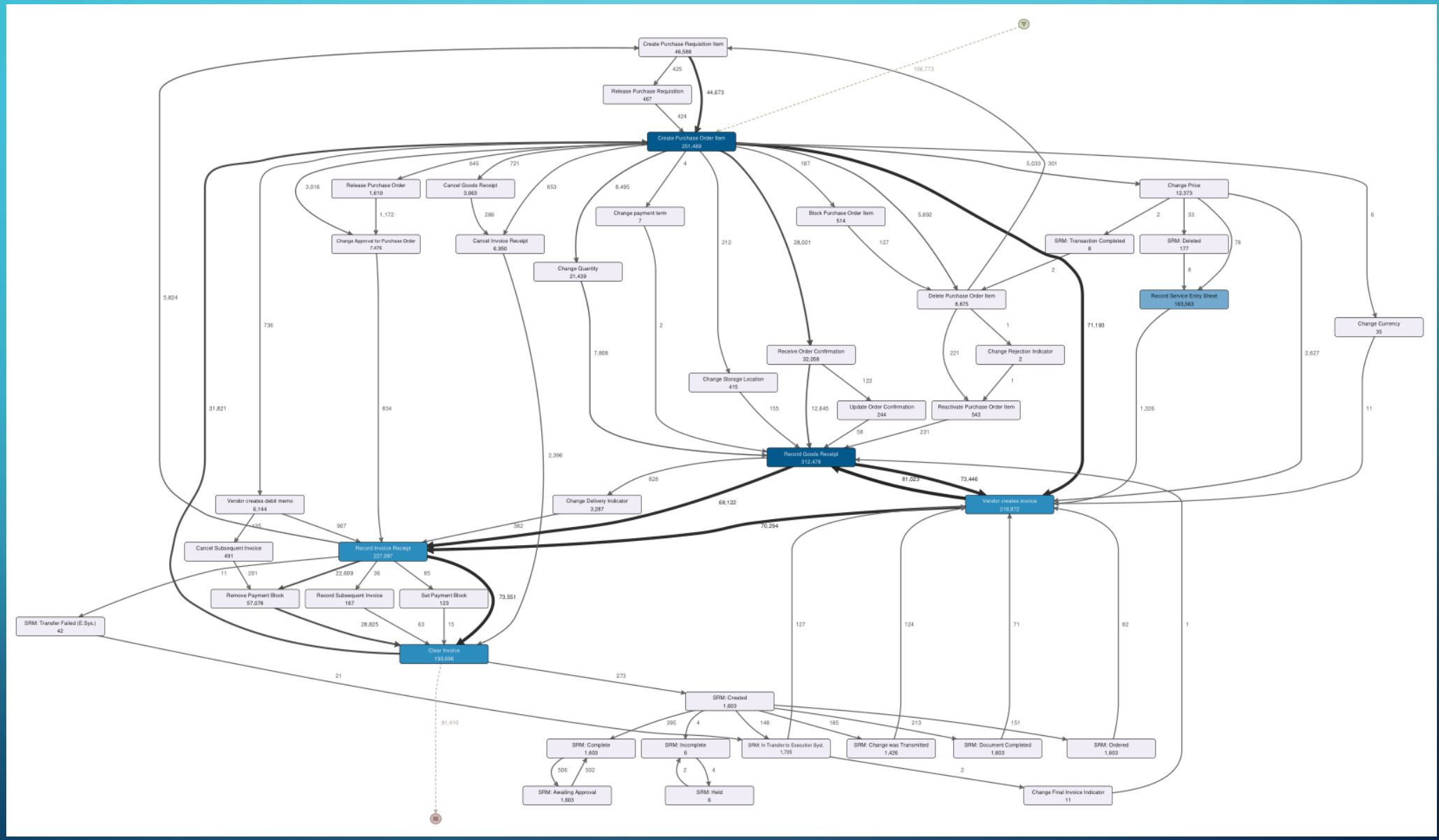
Faruque Ahmed
fahmed.max@gmail.com

Case spend Classification text



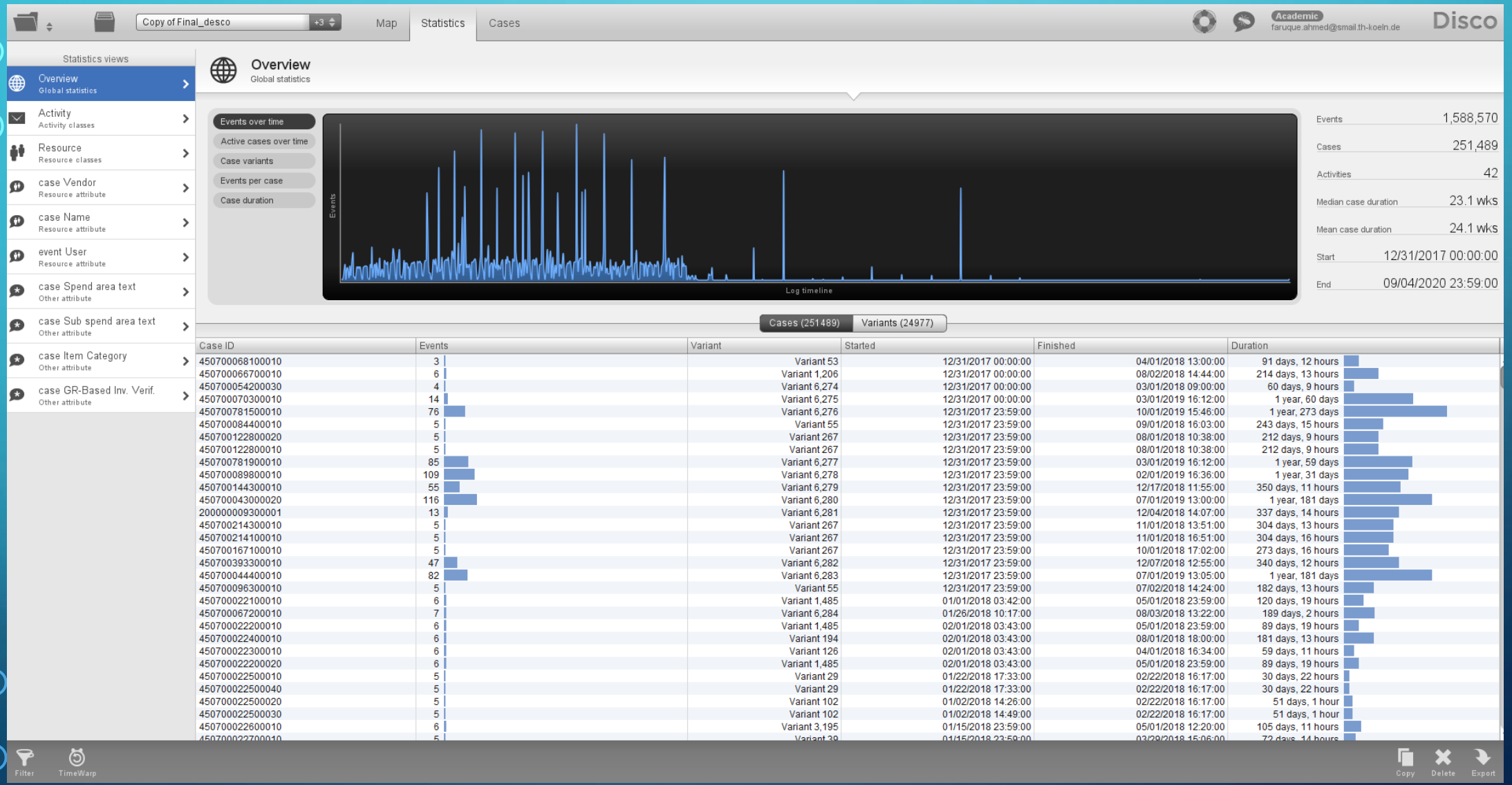
Faruque Ahmed
fahmed.max@gmail.com

Complete Process



Faruque Ahmed
fahmed.max@gmail.com

Dashboard



Case ID	Events	Variant	Started	Finished	Duration
450700068100010	3	Variant 53	12/31/2017 00:00:00	04/01/2018 13:00:00	91 days, 12 hours
450700066700010	6	Variant 1,206	12/31/2017 00:00:00	08/02/2018 14:44:00	214 days, 13 hours
450700054200030	4	Variant 6,274	12/31/2017 00:00:00	03/01/2018 09:00:00	60 days, 9 hours
450700070300010	14	Variant 6,275	12/31/2017 00:00:00	03/01/2019 16:12:00	1 year, 60 days
450700781500010	76	Variant 6,276	12/31/2017 23:59:00	10/01/2019 15:46:00	1 year, 273 days
450700084400010	5	Variant 55	12/31/2017 23:59:00	09/01/2018 16:03:00	243 days, 15 hours
450700122800020	5	Variant 267	12/31/2017 23:59:00	08/01/2018 10:38:00	212 days, 9 hours
450700122800010	5	Variant 267	12/31/2017 23:59:00	08/01/2018 10:38:00	212 days, 9 hours
450700781900010	85	Variant 6,277	12/31/2017 23:59:00	03/01/2019 16:12:00	1 year, 59 days
450700089800010	109	Variant 6,278	12/31/2017 23:59:00	02/01/2019 16:36:00	1 year, 31 days
450700144300010	55	Variant 6,279	12/31/2017 23:59:00	12/17/2018 11:55:00	350 days, 11 hours
450700043000020	116	Variant 6,280	12/31/2017 23:59:00	07/01/2019 13:00:00	1 year, 181 days
200000009300001	13	Variant 6,281	12/31/2017 23:59:00	12/04/2018 14:07:00	337 days, 14 hours
450700214300010	5	Variant 267	12/31/2017 23:59:00	11/01/2018 13:51:00	304 days, 13 hours
450700214100010	5	Variant 267	12/31/2017 23:59:00	11/01/2018 16:51:00	304 days, 16 hours
450700167100010	5	Variant 267	12/31/2017 23:59:00	10/01/2018 17:02:00	273 days, 16 hours
450700393300010	47	Variant 6,282	12/31/2017 23:59:00	12/07/2018 12:55:00	340 days, 12 hours
450700044400010	82	Variant 6,283	12/31/2017 23:59:00	07/01/2019 13:05:00	1 year, 181 days
450700096300010	5	Variant 55	12/31/2017 23:59:00	07/02/2018 14:24:00	182 days, 13 hours
450700022100010	6	Variant 1,485	01/01/2018 03:42:00	05/01/2018 23:59:00	120 days, 19 hours
450700067200010	7	Variant 6,284	01/26/2018 10:17:00	08/03/2018 13:22:00	189 days, 2 hours
450700022200010	6	Variant 1,485	02/01/2018 03:43:00	05/01/2018 23:59:00	89 days, 19 hours
450700022400010	6	Variant 194	02/01/2018 03:43:00	08/01/2018 18:00:00	181 days, 13 hours
450700022300010	6	Variant 126	02/01/2018 03:43:00	04/01/2018 16:34:00	59 days, 11 hours
450700022200020	6	Variant 1,485	02/01/2018 03:43:00	05/01/2018 23:59:00	89 days, 19 hours
450700022500010	5	Variant 29	01/22/2018 17:33:00	02/22/2018 16:17:00	30 days, 22 hours
450700022500040	5	Variant 29	01/22/2018 17:33:00	02/22/2018 16:17:00	30 days, 22 hours
450700022500020	5	Variant 102	01/02/2018 14:26:00	02/22/2018 16:17:00	51 days, 1 hour
450700022500030	5	Variant 102	01/02/2018 14:49:00	02/22/2018 16:17:00	51 days, 1 hour
450700022600010	6	Variant 3,195	01/15/2018 23:59:00	05/01/2018 12:20:00	105 days, 11 hours

Faruque Ahmed
fahmed.max@gmail.com

Thanks

Faruque Ahmed
fahmed.max@gmail.com