Landlord Statement Data Warehouse

Term Project

Data Warehousing & Business Intelligence

1. The Scenario:

I am an agent in a certain letting agency.Now I have tons of landlord statement data for multiple rental properties. The data originates from monthly PDF statements issued by our letting agency. Each PDF contains important financial data for a single property during a specific rental period, including rental income, management fees, repairs, security deposit changes,total(money to landlord) and other related information.

Since there are too many PDF statements, analyze long-term questions can be hard. Therefore, my project decideds to design a Landlord Statement Data Warehouse to convert details from PDFs into structured tables and build a star schema.I hope through this, I can help landlords analyze their property statements and gain a more comprehensive understanding of the properties in our company,at the same time, I can have a knowledge of the earn for our angency.

Project Goals:

- Help landlords analyze their property statements comprehensively
- Gain insights into property performance across the agency portfolio
- Track agency earnings from management fees and services

2. Business Understanding & Questions

Goal: Turn hundreds monthly PDF statements into a central data warehouse so landlords and the agency can analyze property performance over time.
Key questions for my data warehouse to answer:

Rent and Profit Analysis

- How much rent and profit did each property generate over the past (number) years?
- Which properties are the most/least profitable?

Fee Analysis

- What was the total management fee collected by the agency?
- How do fees vary by property or landlord?

Expense and Anomaly Detection

- Which properties had unusually high repair or miscellaneous expenses?
- Are there seasonal patterns in maintenance costs?

3. Source Data Analysis (PDF → Raw Structured Data)

Source: monthly landlord PDF statements, one per property per statement period.

Process in this step:
1.Collect sample PDFs from different properties and months.
2.create CSV table with 42 columns and 134 rows(big enough)
3.Decide how to extract from PDF:
Manual type data/ OCR / machine learning method (as future enhancement).

Now I can see problems from this big table:
Repeating concepts (landlord, property, tenant, lease, charges, agent…) all mixed in this table.
Lots of redundancy (same landlord info repeated for every statement).
Charge columns (rent, management_fee, repair, deposit, misc) are really many rows of just the same type: charge

4. Staging Layer: Clean 1NF Table

make them into 1NF structure, ready for normalization.(1NF requires that all of the keys are defined with no repeated words.)

Choose the PK: statement_id

- Landlord address (such as 45 High Street, Glasgow, G1 1AA) is not atomic, may need to split into street,city, postcode.on the other side, I think maybe it's OK since landlord address is only attached to landlord and don't matter the analysis.
- Charges such as rent/management fee/repair/deposit/misc is repeated in one statement, so I need to add a primary key to it and separate to rows

original

statement_id	rent	rent	management_fee	repair	deposit	misc	note
1542	350	35	35	28/05/1900	0	15	plumbing; credit back

1NF

statement_charge_id	statement_id	charge_type	amount	note
1	1542	rent	350	NULL
2	1542	management_fee	35	NULL
3	1542	repair	149.5	boiler and plumbing
4	1542	misc	15	credit back

Then Check:
1.one row per statement per property.
2.no repeating groups, clear data types, consistent date formats.
3.convert dates to a standard format (YYYY-MM-DD).
4.convert all amounts to a standard currency and numeric type.
5.add a unique statement_id.

5. Normalization(2NF → 3NF / ODS)

Next, normalize to 2NF then to 3NF to remove redundancy and improve data quality.

2NF Normalization: 2NF requires satisfying 1NF and all non-primary attributes are completely depend on primary key.

- Separate landlord from original data(in 1NF table, attributes like landlord_name, email, phone, created_date, is_active depend only on landlord_reg_number, not on statement_id.)

landlord_id	landlord_reg_number	landlord_name	email	phone	address	created_date	is_active
1	7384219/118/10028	James McAllister	james.mcallister@email.com	07700 900101	45 High Street, Glasgow, G1 1AA	2019-06-15	1
2	8642097/210/20001	Sarah Henderson	sarah.henderson@email.com	07700 900102	12 Queen Street, Edinburgh, EH2 1AB	2020-01-10	1
3	1259924/395/06072	Robert Campbell	robert.campbell@email.com	07700 900103	78 Main Road, Paisley, PA1 2CD	2018-03-22	1

- Also since a landlord can have many bank accounts, I also separate it into one table:

bank_account_id	landlord_id	account_name	bank_name	account_number	sort_code	is_primary	created_date
1	1	James McAllister	Bank of Scotland	12345678	80-11-22	1	2019-06-15
2	2	Sarah Henderson	Royal Bank of Scotland	23456789	83-22-33	1	2020-01-10
3	3	Robert Campbell	Barclays	34567890	20-33-44	1	2018-03-22
4	3	R Campbell Savings	Nationwide	45678901	07-01-16	0	2020-05-15

- Separate property from original data(in the flat table, property_alias, bedrooms, postcode, property_type_name all depend on property, not statement.)
- Separate tenant (Tenant personal info depends on tenant, not the particular statement.)

tenant_id	first_name	last_name	email	phone	emergency_contact	created_date
1	Michael	Brown	m.brown@email.com	07712 345678	Jane Brown 07700 111222	2022-04-15
2	Emma	Wilson	e.wilson@email.com	07723 456789	Mark Wilson 07700 222333	2022-12-01
3	David	Taylor	d.taylor@email.com	07734 567890	Susan Taylor 07700 333444	2020-09-10
4	Lisa	Anderson	l.anderson@email.com	07745 678901	Tom Anderson 07700 444555	2021-06-20
5	John	Murray	j.murray@email.com	07756 789012	Mary Murray 07700 555666	2023-01-15
6	Karen	Stewart	k.stewart@email.com	07767 890123	Paul Stewart 07700 666777	2024-04-01

- Separate lease(Lease dates & monthly rent depend on the lease, not directly on each statement.)
- Separate property and agent assignment(M:M):A property can be managed by different agents over time; an agent can manage many properties.Instead of repeating agent columns on each property, I create a table:property assignment.

agent_id	first_name	last_name	email	phone	role	hire_date	is_active
1	Emma	Thompson	emma.t@agency.com	0141 332 1001	Property Manager	2018-03-15	1
2	David	Mitchell	david.m@agency.com	0141 332 1002	Property Manager	2019-07-01	1
3	Sophie	Clark	sophie.c@agency.com	0141 332 1003	Senior Manager	2015-01-10	1
4	Alan	Reid	alan.r@agency.com	0141 332 1004	Administrator	2020-06-01	1

assignment_id	property_id	staff_id	assignment_date	end_date	is_primary
1	1	1	2019-06-15		1
2	2	1	2020-01-15		1
3	3	2	2018-03-25		1
4	1	3	2022-01-01		0
5	2	3	2022-01-01		0

- Separate contractor since they are only depends on contractor

contractor_id	company_name	contact_name	email	phone	service_type	is_approved
1	QuickFix Plumbing	John Smith	info@quickfixplumbing.co.uk	0141 445 1111	Plumbing	1
2	Spark Electric	Mary Johnson	mary@sparkelectric.co.uk	0141 445 2222	Electrical	1
3	SafeGas Services	Bob Williams	bob@safegasservices.co.uk	0141 445 3333	Gas/Heating	1
4	CleanHome Services	Sarah Davis	sarah@cleanhome.co.uk	0141 445 4444	Cleaning	1
5	SecureHomes Locksmiths	Mike Thompson	mike@securehomes.co.uk	0141 445 5555	Locksmith	1
6	GreenThumb Gardens	Claire Wilson	claire@greenthumb.co.uk	0141 445 6666	Landscaping	1

3NF Normalization:

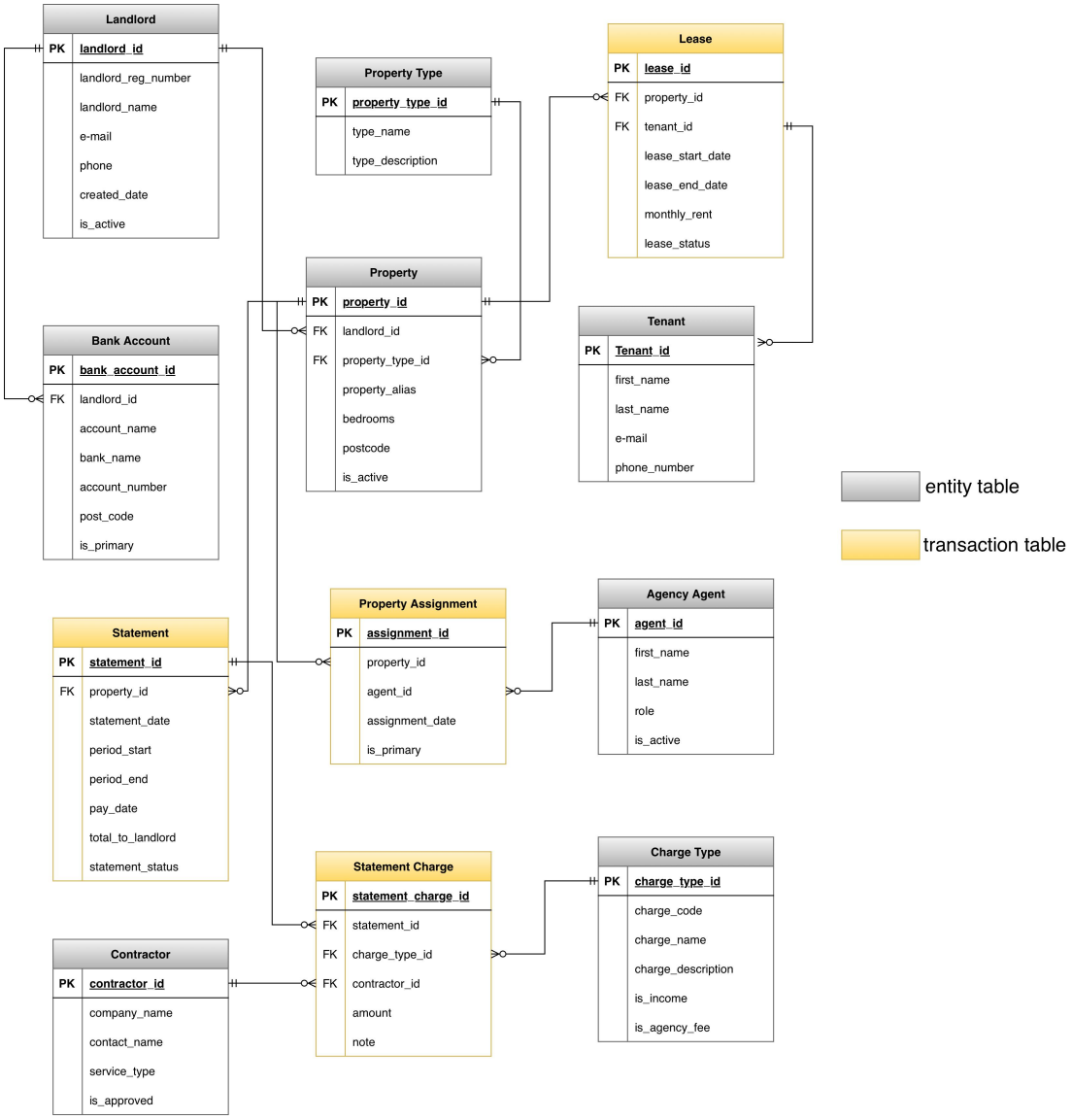
3NF requires satisfying 2NF and non-primary attributes cannot depend on other non-primary attributes. So normalize to:

ENTITY TABLES:

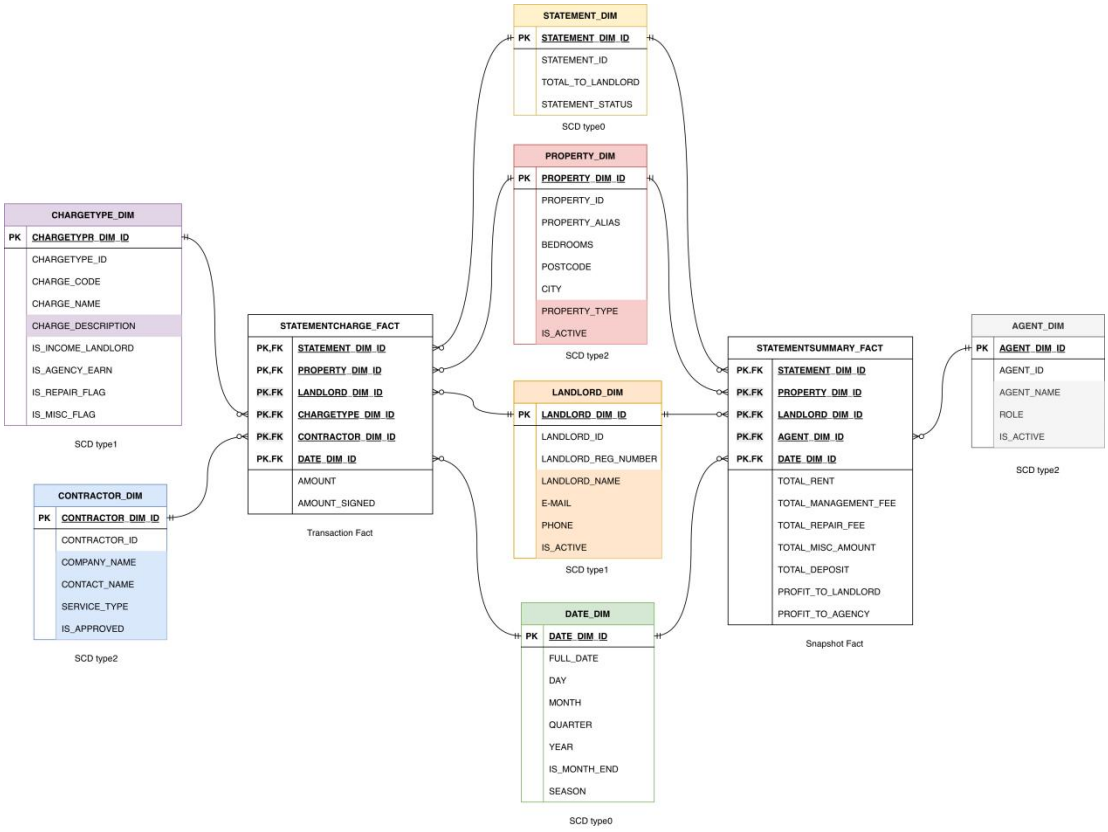
1. Landlord - Property owners
2. Bank Account - Landlord's bank details
3. Property Type - Flat, Terraced, etc.
4. Property - Individual properties
5. Tenant - Renters
6. Agency Agent - Staff of agency
7. Contractor - Service providers
8. Charge Type - Types of charges (rent, repairs, fees)

RELATIONSHIP/TRANSACTION TABLES:

9. Lease - Links Property to Tenant
10. Property Assignment - Links Property to Agent
11. Statement - Monthly statements
12. Statement Charge - Individual line items on statement



Dimensional Modeling: Star Schema Design


STATEMENT_DIM
PK STATEMENT_DIM_ID
STATEMENT_ID
TOTAL_TO_LANDLORD
STATEMENT_STATUS
SCD type0
PROPERTY_DIM
PK PROPERTY_DIM_ID
PROPERTY_ID
PROPERTY_ALIAS
BEDROOMS
POSTCODE
CITY
PROPERTY_TYPE
IS_ACTIVE
SCD type2
CHARGETYPE_DIM
PK CHARGETYPR_DIM_ID
CHARGETYPE_ID
CHARGE_CODE
CHARGE_NAME
CHARGE_DESCRIPTION
IS_INCOME_LANDLORD
IS_AGENCY_EARN
IS_REPAIR_FLAG
IS_MISC_FLAG
SCD type1
STATEMENTCHARGE_FACT
PK,FK STATEMENT_DIM_ID
PK,FK PROPERTY_DIM_ID
PK,FK LANDLORD_DIM_ID
PK,FK CHARGETYPE_DIM_ID
PK,FK CONTRACTOR_DIM_ID
PK,FK DATE_DIM_ID
AMOUNT
AMOUNT_SIGNED
Transaction Fact
LANDLORD_DIM
PK LANDLORD_DIM_ID
LANDLORD_ID
LANDLORD_REG_NUMBER
LANDLORD_NAME
E-MAIL
PHONE
IS_ACTIVE
SCD type1
STATEMENTSUMMARY_FACT
PK,FK STATEMENT_DIM_ID
PK,FK PROPERTY_DIM_ID
PK,FK LANDLORD_DIM_ID
PK,FK AGENT_DIM_ID
PK,FK DATE_DIM_ID
TOTAL_RENT
TOTAL_MANAGEMENT_FEE
TOTAL_REPAIR_FEE
TOTAL_MISC_AMOUNT
TOTAL_DEPOSIT
PROFIT_TO_LANDLORD
PROFIT_TO_AGENCY
Snapshot Fact
AGENT_DIM
PK AGENT_DIM_ID
AGENT_ID
AGENT_NAME
ROLE
IS_ACTIVE
SCD type2
CONTRACTOR_DIM
PK CONTRACTOR_DIM_ID
CONTRACTOR_ID
COMPANY_NAME
CONTACT_NAME
SERVICE_TYPE
IS_APPROVED
SCD type2
DATE_DIM
PK DATE_DIM_ID
FULL_DATE
DAY
MONTH
QUARTER
YEAR
IS_MONTH_END
SEASON
SCD type0